National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, tests publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30.

NL-339 (r. 88/04)

Canada

The University of Alberta

Representing and Reasoning with Probabilistic Knowledge

by

© Fahiem Bacchus

A thesis
submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree
of Doctor of Philosophy

Department of Computing Science

Edmonton, Alberta
Fall 1988

Permission has been granted to the National Library of Canada to microfilm this thesis and to lend or sell copies of the film.

The author (copyright owner) has reserved other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without his/her written permission.

L'autorisation a été accordée à la Bibliothèque nationale du Canada de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur (titulaire du droit d'auteur) se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation écrite.

ISBN 0-315-45680-9

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Fahiem Bacchus
TITLE OF THESIS: Representing and Reasoning with Probabilistic Knowledge

DEGREE: Doctor of Philosophy
YEAR THIS DEGREE GRANTED: 1988

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis not extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed)

Permanent Address:
10935–40a Ave.
Edmonton, Alberta,
Canada, T6J–0T1.

Date: July 7th 88

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled **Representing and Reasoning with Probabilistic Knowledge** submitted by **Fahiem Bacchus** in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

L K Schubert
(Supervisor)

Renée J. Elio

Mohan Matthew

R. Goebel

Date: July 4th 1988

For my parents

Abstract

This thesis presents a logical formalism for representing and reasoning with probabilistic knowledge. The formalism differs from previous efforts in this area in a number of ways. Most previous work has investigated ways of assigning probabilities to the sentences of a logical language. Such an assignment fails to capture an important class of probabilistic assertions, empirical generalizations. Such generalizations are particularly important for AI, since they can be accumulated through experience with the world. Thus, they offer the possibility of reasoning in very general domains, domains where no experts are available to gather subjective probabilities from.

A logic is developed which can represent these empirical generalizations. Reasoning can be performed through a proof theory which is shown to be sound and complete. Furthermore, the logic can represent and reason with a very general set assertions, including many non-numeric assertions. This also is important for AI as numbers are usually not available.

The logic makes it clear that there is an essential difference between em-

pirical, or statistical, probabilities and probabilities assigned to sentences, e.g., subjective probabilities. The second part of the formalism is an inductive mechanism for assigning degrees of belief to sentences based on the empirical generalizations expressed in the logic. These degrees of belief have a strong advantage over subjective probabilities: they are founded on objective statistical knowledge about the world. Furthermore, the mechanism of assigning degrees of belief gives a natural answer to the question "Where do the probabilities come from:" they come from our experience with the world.

The two parts of the formalism offer combined, interacting, but still clearly seperated, plausible inductive inference and sound deductive inference.

Acknowledgements

I am pleased to acknowledge the contributions of my teacher Len Schubert. His AI course got me interested in the area, and he suggested this line of research. His insightful comments, as well as his keen intuitions, played a major role in the eventual generality of the formalism. Thanks also to my external examiner Joe Halpern, who read my thesis very carefully and gave many important and useful comments, comments which made me far more confident as to its mathematical soundness. Henry Kyburg provided me with encouragement and funded a very rewarding semester long visit to the University of Rochester. Teddy Seidenfeld introduced me to his joint work with Schervish and Kadane on non-sigma-additive probabilities. Jeff Pelletier gave me some very useful pointers into the philosophical literature. Randy Goebel provided many useful comments. And Rene Elio and Mohan Mathan took the time to read my thesis. Thanks to all.

Various graduate students at the University of Alberta and at the University of Rochester made my Ph.D. experience far more enjoyable. An incomplete list includes, Ambrish Mathur, Cao-an Wang, Bonita Wong, George Ferguson, Scott Goodwin, Chung Hee Hwang, Stephanie Miller, Franco Carlacci, Jay Webber, Alice Kyburg, and Leo Hartman. Special thanks to Nicola Ferrier and Josh Tenenberg.

Family and other friends away from the university deserve a special acknowledgement for allowing me to escape from my studies when I needed to. Thanks to Narry, Tom and Zeeda, Judy, Teddy, my other Mom and Dad, the kidlings, Zoie, Kosta, Maria and George, Barb and Tim, Najm and Ruby, David and many others.

And Liz was there.

Contents

1 Introduction 1

1.1 Logic and Knowledge Representation 1

1.1.1 First Order Logic 4

1.2 Probabilities 6

1.2.1 Types of Probabilities 8

1.3 The Contribution of This Work 10

1.4 Non-Universal Generalizations 13

1.5 Outline of the Presentation 19

2 Introduction to the Formalism **21**

2.1 The logic **Lp** 21

2.1.1 Types of Statistical Knowledge 21

2.1.2 The Field of Numbers 25

2.1.3 Probabilities over the Domain of Discourse 27

2.1.4 Non-Standard Probabilities 32

2.1.5 Related Work 35

2.2 Belief Formation 39

2.2.1 The Inductive Assumption of Randomization 40

2.2.2 Conflicting Knowledge 42

2.2.3 Kyburg's Work 44

2.3 A Simple Rational Agent 45

3 The Logic Lp 48

3.1 Symbols 49

3.2 Formulas 50

3.3 Semantic Model 52

3.4 The Effect of the Coherence Constraints 59

3.5 Semantics of Formulas 62

3.6 Examples of Knowledge Representable in **Lp** 65

4 Deductive Proof Theory 68

4.1 Substitution 71

4.2 Proof Theory 86

4.2.1 Axioms and Rules of Inference 86

4.2.2 Deductions 90

4.2.3 Provable Equivalence 92

4.2.4 Maximal Consistency 97

4.2.5 Soundness and Completeness of the Proof Theory . 100

4.3 Properties of the Probability Terms 112
4.4 Examples of Reasoning with the Statistical Knowledge . . . 116

5 Belief Formation **121**
5.1 Belief Formation 123
5.1.1 Inductive Evaluation Function 124
5.1.2 Preference Criterion 127
5.1.3 Properties of the Mechanism 130
5.2 Semantic Justification 132
5.3 Belief Formation—Examples 133

6 An Inheritance Reasoner **142**
6.1 The Nature of Inheritance Systems 144
6.2 Heterogeneous vs. Homogeneous Inheritance Systems 148
6.3 Representing Statements of Typicality With Probabilities . 151
6.4 The Inheritance Graph 155
6.5 Inferences in the Inheritance Net 158
6.6 The Relation to Belief Formation 161
6.6.1 On the Independence Assumptions Underlying Inheritance Reasoners 164
6.7 Behavior of the Reasoner 166
6.8 Touretzky et al.'s Design Space 170
6.8.1 Skepticism vs. Credulity 171

6.8.2 Upwards vs. Downwards Reasoners 175

6.8.3 On-Path vs. Off-Path Preemption 178

6.9 Extending the Graph's Expressiveness 180

7 Conclusion **183**

7.1 What has Been Accomplished 183

7.2 Future Research 186

7.2.1 Extensions to Belief Formation 187

7.2.2 Diagnosis from Statistical Principles 192

7.2.3 Learning, or Induction, from Ground Facts 195

List of Figures

1 The Graph Encoding 157

2 Redundant Information 167

3 Ambiguous Information 168

4 (M. Ginsberg) Is Nixon Politically motivated? 169

5 Tweety is probably not a penguin 170

6 Two examples of cascaded ambiguities 173

7 Nodes A and B are coupled 176

8 Opportunism 177

9 Two examples of On-path vs. Off-path preemption 179

Chapter 1

Introduction

1.1 Logic and Knowledge Representation

It is well accepted in AI that the performance of a task requires an explicit body of relevant knowledge. As a result most research in AI has used a knowledge-based approach. A mechanism for representing knowledge in a computer model or program is a requirement of this approach.

A natural consequence of this has been a desire to develop *general* mechanisms for knowledge representation. It is clear that a truly intelligent entity is not a specialist: given a sufficient diversity of knowledge, an intelligent entity is capable of performing a wide variety of tasks. Hence, to the extent that the eventual aim of AI is to develop such an intelligence, it will be necessary to represent a variety of knowledge in the same program.

Furthermore, even less ambitious goals, such as the construction of natural language understanding systems (or maybe as ambitious, if one agrees with Turing), require a large variety of general knowledge. Each time a representation scheme is developed to handle a particular type of knowledge a recurring set of problems arise. A general mechanism allows these problems to be dealt with once, rather than continually.

The problems which arise come from the need to specify the meaning of the objects in the representation scheme. Say knowledge is represented in a AI program through the use of some particular data structures, that inference is performed through the algorithmic manipulation of these structures, and that output is generated from the modified data structures. Then unless there is a clear meaning attached to these data structures it becomes impossible to understand, or validate, the outputs of the program. Exactly what knowledge is contained in the program, exactly what the output represents (i.e., what new knowledge it encodes), and how the output is to be justified will all be uncertain.

Formal logics possess the important property of transparency of meaning. That is, once we have assigned a meaning to the non-logical symbols in the logic, each expression, no matter how complex, will be given a well defined meaning. This meaning is imparted through the logic's formal semantics, and is built up recursively in a one-to-one manner with the construction of the expression. Since expressions can be built up recursively logics can be very general mechanisms for knowledge representation without sacrificing well defined meaning.

Furthermore, most logics also have a well specified theory of inference. For example, deduction in first order logic. These inference procedures give formal guaranties, or justifications, to their conclusions. For example, deduction guaranties that if the knowledge used is true then the conclusions generated will also be true.

It is important to note that AI programs need not use expressions of a logic as their internal representation, nor do they need to use the inference procedure specified by the logic. Indeed, it is well known that even for simple propositional logics this inference procedure is NP-complete, while for ordinary first order logic it is undecidable. However, this does not preclude analyzing the program by relating its internal structures and its inference algorithms to a logic. The logic will provide the tools necessary for a rigorous analysis of the program. Nor does the general intractability of inference in logics mean that AI programs cannot use logics directly, with expressions in a logic as its internal representation and implementing the logic's inference procedure. Automated theorem provers have been able prove many useful theorems even though they are implementing an undecidable inference procedure. Nor are these two uses of logic incompatible. In fact, work by Schubert et al. [73] has demonstrated that special representations and algorithms which are formally related to expressions and deduction in first order logic can profitably co-exist in a program which also uses logic directly.

1.1.1 First Order Logic

Ordinary first order logic has been the primary formal logic used in AI for knowledge representation. It has the advantage of having a mechanizable inference procedure which gives its conclusions a very strong justification. The inference procedure is, of course, deduction, and the justification is that if the knowledge used (i.e., the knowledge stored in the program) is true then the conclusions *must* also be true. For example, starting with the knowledge that penguins are birds and that Opus is a penguin, deduction can produce the conclusion that Opus is a bird. This conclusion is necessarily true if the premises are true.

There is a large set of knowledge which is true simply in virtue of terminological conventions, e.g., bachelors are unmarried men, penguins are birds. Any deductive consequences of such knowledge must also be accepted as being true. Deduction as a mechanism of inference is not, however, limited to such analytically true statements. It can also be applied to statements whose truth values are not known but which are believed to be true. Most scientific knowledge, for example, falls into this category. Having accumulated sufficient evidence to believe that Newton's laws of motion yield good approximations at sub-light velocities, a scientist can deduce from these laws (premises) certain conclusions about the motion of the planets, such as, "planets describe an elliptical orbit," which he can accept with equal conviction. Deduction can even be applied to statements not yet believed due to insufficient evidence, e.g., to explore the consequences of a new hypothesis, or in a backwards manner, e.g., to generate possible theories which explain the known observations (e.g., Poole [62], Morgan [51]).

First order logic, although powerful, cannot easily represent all of the knowledge that people use.[1] Nor is its deductive proof theory sufficient for all of the reasoning performed. One of the shortcomings lies in its inability to deal in a reasonable manner with plausible inferences, that is, inferences which are not always true given the premises (i.e., not deductive conclusions), but which are plausible, or probable, and can be believed at varying levels of certainty.

---

[1] A great deal of knowledge can be expressed in first order logic through the use of devices like set theory. With the axioms of set theory in the language much of mathematics can be built up and many complex concepts thus expressed. However, the addition of the axioms of set theory makes deduction in the logic so complex that any sort of automation becomes very difficult.

1.2 Probabilities

Many ordinary inferences have a tentative character. For example, if we have the knowledge that most birds fly and we are told that Tweety is a bird, a plausible inference is that Tweety can fly. Deduction is not capable of this kind of inference—in this case the truth of the premises does not imply the truth of the conclusion. There may be any number of reasons why Tweety might not be able to fly, e.g., he might be a penguin, he might have a broken wing, he might be an ostrich, etc. However, irrespective of these negative possibilities, it is clear that the ability to make such inferences is essential for the successful performance of various tasks. In many situations it is necessary to act, to draw some conclusion, despite the fact that no conclusions are possible through deduction.

It is natural to look to probabilities for a means of modeling these types of inferences. Probabilities were originally developed with such a use in mind. Specifically, they were developed as a rational guide to action in gambling, where, for example, one does not know what the outcome of the next toss of the die will be, and furthermore, one has insufficient knowledge to deduce the outcome. Since they were first developed, probabilities have been the subject of much study, which has resulted in a simple and well understood formalization (Kolmogoroff [35]). Probabilities have also been used by philosophers in their studies of non-deductive, or inductive, inferences. Such work has produced some very strong reasons for the use of probabilities. These results are normative results, i.e., they show that in order to insure certain criteria of rationality probabilities are the only possible formalism when one wishes to hold graded levels of belief.

There are two major normative results, the results of Cox and the results of De Finetti and others. Cox [11][2] developed a set of simple criteria intuitively required of any measure which represents a degree of belief in an assertion. He then proceeded to prove that any measure which satisfied these criteria must be a probability function. De Finetti took a decision theoretic approach to probabilities, viewing probabilities as being a guide to action. He was able to prove that any measure over a propositional lattice of sentences used to guide decisions must be a probability function (see [8, chapters 3 and 7]); any other function will lead to certain loss is some situations. These results were extended, independently, by Shimony [74], Kemeny [34], and Sherman [42], who showed that any function which given some evidence assigns a degree of belief to a hypothesis must be a conditional probability function, if that degree of belief is to be used to guide action.

The fact that these global justifications exist is a very important reason for looking to probabilities. As mentioned above, deduction gives a very strong justification to its conclusions, but that justification is, in many situations, too strong: it cannot be hoped that any mechanism for *plausible* inference can be justified in such a strong manner. At the same time it is important that such a mechanism be given some sort of global justification, i.e., a justification which is independent of context. Citation of specific examples of intuitive behaviour can never be a substitute for such justifications.

---

[2] Also see Horvitz et al. [31] for a discussion of the implications of Cox's work to AI, and Aleliunas [2] for a generalization of Cox's results to probability functions whose range is less structured that the reals.

1.2.1 Types of Probabilities

Although the formal model of probabilities is well defined, its meaning, or interpretation, remains a contentious issue. There are three major interpretations of probabilities: the empirical, the logical, and the subjective (see Kyburg [37, Chapter 1] and for more detail Kyburg [38]). The empirical interpretation is the oldest of the three, having been explicitly proposed at least as early as 1866 by John Venn [80]; it has been adopted by many writers including Neyman [52], Reichenbach [64], Salmon [68], von Mises [81], and Popper [63]. The empirical interpretation takes probability statements as representing statistical truths about the world. In this view, probability statements are objective statements about the world, and their truth or falsity has nothing to do with a person's opinions, or any body of evidence, but only with the state of the world. This view of probabilities can be contrasted with the logical interpretation (Carnap [9], Hintikka [29]), which views probabilities as being a relationship between a body of evidence and an assertion, both expressed in a fixed logical language, the relationship being determined by the rules of the language. Finally, the subjective, or personalistic, interpretation (De Finetti [14], Savage [70]), views probabilities as being degrees of belief, or propensity to take action, held by a particular person at a particular time. These degrees of belief need not have anything to do with reality; they are purely subjective.

When we examine the different types of knowledge we wish to represent in AI programs it is found that a lot of it is in the form of statistical claims about the world. These range from imprecisely quantified generalizations, like "Most birds fly," or, "People with a runny nose usually have a cold," to more precisely quantified statistical statements, like those found in various expert systems which deal with uncertainty (e.g., the MYCIN system [75]). The interesting thing about such statistical claims is that they play a different role in a knowledge based system than the plausible inferences drawn from them, even though they are both in some sense probabilistic. These statistical claims are used in a manner similar to objective logical knowledge. For example, when making the plausible inference "Tweety can fly" the statistical claim "Most birds fly" is being used in the same manner as the knowledge "Tweety is a bird." That is, both are assumed to be objectively true pieces of knowledge. The plausible inference is, on the other hand, a conjecture believed at some level of certainty. The statistical claim is being interpreted as a statistical truth about the world, i.e., as an empirical probability, whereas the plausible inference is being interpreted as an assertion held to some degree of belief, i.e., a subjective or logical probability.

1.3 The Contribution of This Work

The major contribution of this work is to provide a formal logic capable of representing a wide variety of statistical claims, through an empirical probabilistic component in the semantics. The logic is an extension of ordinary first order logic, and its development is complete; that is, not only are the syntax and semantics specified, but also a sound and complete deductive proof theory is provided. This proof theory is capable of reasoning with statistical knowledge, as well as with sentences of first order logic.

Since statistical knowledge is represented by statements of empirical probability, it has a logic of its own, imparted by the logic of probabilities. For example, from the generalization "Most birds fly" along with the knowledge "Penguins are birds" and "Penguins do not fly," it is possible to deduce that "Most birds are not Penguins." When more precisely quantified knowledge is present, all of the power of Bayesian analysis is available, including causal reasoning, and weighing of evidence. In fact, since the proof theory is complete it logically subsumes most of the deductive probabilistic reasoning systems (i.e., systems based on the axioms of probability) previously developed.

The second contribution is a general mechanism of plausible inference, called belief formation, which can use the statistical knowledge to generate plausible conclusions about particular cases. For example, the mechanism of belief formation is capable of making the inference that Tweety can probably fly, given the knowledge that Tweety is a bird and that most birds fly.

Belief formation generates a graded conclusion in that it assigns a degree of belief to the conclusion, a degree which ranges from zero to one and which obeys the laws of probability. The degree of belief is similar to an assignment of a subjective probability to the conclusion, and it can be used to guide action in the face of uncertainty. Since these measures of belief obey the laws of probability, the normative results mentioned above can be used to justify their use. The major difference between this degree of belief and a subjective probability, as defined by De Finetti and Savage, is that a subjective probability is based only on a person's subjective opinions and *need* have no basis in reality. If subjective probabilities are used to guide action, then these probabilities can be considered to be 'correct' to the extent that one's actions are rational.[3] The degrees of belief generated through belief formation, on the other hand, are not based on subjective opinions; they are based on knowledge which is considered to be representative of certain statistical truths about the world.

One feature of the plausible conclusions which can be drawn from the statistical knowledge is that they display non-monotonic behaviour. That is, a conclusion sanctioned by the system may no longer be sanctioned when new facts are *added* to the system.[4] In the simplest case a plausible conclusion may later be learned to be false, i.e., its negation may be added as a new fact. A slightly more complex case occurs when new evidence is added which changes your degree of belief in a previous plausible conclusion. For example, you may conclude that it is probable that Tweety can fly, but upon learning some new information, e.g., that Tweety has had his wings clipped, you may be forced to retract the previous conclusion, and conclude instead that it is unlikely that Tweety can fly. That is, the degree of belief in Tweety flying has changed from some high value to some low value. In a still more complicated case, new information may be added which makes the statistical knowledge upon which the plausible inference is founded inapplicable. For example, you may conclude upon observing that John has a runny nose that John probably has a cold, but later you may learn that he suffers from allergies, which makes you question the applicability of your statistical knowledge about runny noses and colds.

---

[3] This could be the justification for the use of subjective probabilities by some researchers (e.g., Duda et al. [17]) in expert systems. That is, the experts' subjective opinions are considered to be an accurate reflection of reality.

[4] Hempel [28] calls this behaviour the "Ambiguity of Statistical Systematization" and gives a lengthy discussion on why it occurs.

A reasoning system which consists of first order logic along with deduction is, on the other hand, purely monotonic, i.e., a conclusion will never be invalidated by the addition of new knowledge. This is a result of the strong justification imparted by deduction to its conclusions. As long as the premises used in the deduction remain in the system the conclusion will remain a valid deduction, regardless of what new knowledge is added to the system.

1.4 Non-Universal Generalizations

There has been a extensive amount of work in AI addressing the problem of representing and reasoning with non-universal generalizations, i.e., generalizations which admit exceptions. Notably, various schemes have been developed for non-monotonic reasoning which address, among other things, this problem (see, e.g., the articles in [4]). The system constructed here is capable of dealing with a large set of these generalizations, in particular, with those generalizations which can be given a statistical interpretation, e.g., empirical generalizations like "People with a runny nose usually have a cold", "It almost never rains in Southern California in the summer", "It's usually very cold in January in Edmonton", "Most birds can fly", "Milk is usually kept in a fridge."

There are a number of non-universal generalizations which seem to have more than a statistical interpretation. One common type are those generalizations which take the form of generic sentences in English, for example, the sentence "Cats are stealthy." This sentence seems to have an intentional meaning as well as an extensional meaning. The intentional meaning says that this is some sort of identifying property of cats, i.e., a property which will be preserved across possible worlds. The extensional meaning may be more statistical, e.g., in most situations (in this world) cats are stealthy, (see Schubert and Pelletier [60] for a discussion of these matters). The intentional part of the meaning of this sentence is beyond the capabilities of the formalism developed here, hence it cannot capture the full meaning of such statements.

What is claimed here is that the set of non-universal generalizations which can be given a statistical interpretation represents a large and important set of knowledge which is necessary for the performance of a number of different tasks in AI.

The popular example of "Most birds fly" can be used to compare some of the other work in non-monotonic reasoning to the work presented here. It should also be noted, however, that the developers of most non-monotonic reasoning systems have much more in mind than just the modeling of non-universal generalizations.[5] In the non-monotonic system of default logic constructed by Reiter [65] this generalization could be represented by a default rule,

$$\frac{\mathrm{Bird}(x)\colon \mathbf{M}\mathrm{Fly}(x)}{\mathrm{Fly}(x)},$$

where the meta-logical operator **M** indicates consistency. In the system of non-monotonic logic of McDermott and Doyle [49] this example would be represented using a formula containing a sentential consistency operator:

$$\forall x((\mathrm{Bird}(x) \wedge \mathbf{M}\mathrm{Fly}(x)) \rightarrow \mathrm{Fly}(x))$$

Intuitively the meanings are similar. Both formulas assert that the conclusion that x can fly can be drawn if x is a bird and if it cannot be proved that x cannot fly (i.e., $\mathrm{Fly}(x)$ is consistent with what is known). McCarthy's circumscription [47] is somewhat different. In his formalism one circumscribes an 'abnormality' predicate to minimize the number of non-flying birds.

Reiter's and McCarthy's systems are meta-language constructions built upon ordinary first order logic. The meta-language constructions encode the non-universal generalizations, and are used as rules which can add certain sentences to an underlying first order theory. However, the only inference mechanism provided is first order deduction; hence, there is no mechanism for reasoning with the generalizations, since the meta-language constructions are outside of first order logic. Thus, even outright contradictions can be present in their systems, e.g., the assertions that "penguins are birds" and "typically penguins are not birds," can both be present without contradiction.[6]

---

[5] McCarthy [46] presents an ambitious list of conjectured uses.

Delgrande [16] has addressed this particular problem, and has constructed a conditional logic, with possible world semantics, which is capable of doing some reasoning with the non-universal generalizations (defaults). It is particularly interesting to note that his system consists of two parts, as does the work presented here. He finds that his deductive modal logic, while capable of representing and reasoning with the non-universal generalizations, is incapable of making the kinds of plausible inferences sanctioned by these generalizations. That is, his logic can represent and reason with statements like "most birds can fly," but is incapable of making the plausible inference "Tweety can fly" when it is known that Tweety is a bird. In order to make these inferences he has to use certain inductive assumptions, as is required by the system developed here. This is not surprising. Since these inferences are only plausible, not certain, they must be based on some sort of inductive, i.e., non-deductive, rules of inference. There are two inductive assumptions used in his system, an assumption of normality, which is similar to McCarthy's circumscription of abnormality predicates, and an assumption of minimal relevance, which is similar to a probabilistic assumption of independence.

---

[6] In McDermott and Doyle's system the consistency operator is in the object language. However, they provide no semantics for this operator. Hence, there is no way of relating the truth of $M\beta$ to the truth of β (Davis [12]). Thus, their system, as they present it, also allows this contradiction.

The probabilistic approach developed in this thesis has a number of advantages over these formalisms for dealing with empirical generalizations. Also, even though the developers of most non-monotonic reasoning systems claim that their formalisms deal with more than just empirical generalizations, it is not clear just how general their formalisms really are, since the intuitions behind these formalisms[7] seem to be of limited applicability.

The first advantage is representational power. These formalisms are limited to "almost always true" statements. This is a result of the ungraded nature of the plausible conclusions generated. The probabilistic approach is capable of expressing, and differentiating between, generalizations which range from almost always true to just as often false as true. Also, due to a key innovation, arithmetic relationships between these generalizations can be expressed. For example, the statement "It is more likely that a politician is a lawyer than an engineer" can be expressed in the probabilistic formalism developed here. Furthermore, it can be expressed without any commitment to the likelihood of either case, i.e., without any commitment either to

---

[7] That is, that the generalizations should apply unless they can be proved to be inapplicable [Reiter, McDermott and Doyle], or, that there are a minimal number of abnormalities [McCarthy, Delgrande].

"most politicians are lawyers" or "most politicians are not engineers."

Another advantage is that these formalisms have, to my knowledge, no formal justifications like those which have been demonstrated for probabilistic degrees of belief.[8] This is a serious limitation, since there are a wide variety of conjectured uses for these formalisms. It is clear that citing specific examples where the theory appears to work is no substitute for global justifications, especially when the range of claimed uses is so wide as to preclude the possibility of empirical testing.

Finally, the probabilistic approach presented is able to reason with the empirical generalizations in a much more general manner that Delgrande's conditional logic. Again, this is a result of the ungraded conclusions generated by his system. The fact that the conclusions are ungraded makes it impossible to weigh evidence in any sophisticated manner, as, for example, is allowed by Bayesian analysis. Furthermore, the inductive assumption of minimum relevance used by Delgrande has a formal analogue in probability theory, as independence. The logic developed here is capable of explicitly representing independence assertions, and is also capable of reasoning with them; hence, this type of assumption can be asserted at a much finer grain. That is, we are not stuck with a global assumption of independence; instead, the exact limits of such an assumption can be represented. The other non-monotonic formalisms are not capable of any sort of reasoning with non-universal generalizations.

---

[8] One problem which results from the lack of any formal justification is "A Clash of Intuitions," Touretzky [78].

1.5 Outline of the Presentation

The next chapter provides the motivation for the formalism developed in the thesis. The logic used to represent statistical knowledge is introduced along with some of the considerations which influenced its final form. The logic is a type of probability logic. Hence, the chapter also contains a discussion of, and comparison with, previous work in probability logics. The mechanism of belief formation is motivated by first explaining the need for an inductive inference mechanism and then the intuition upon which it is based. A simple model of a rational agent is presented, which helps to clarify the role of the two parts of the formalism. A closely related system of inductive inference, previously constructed by Kyburg [37], is mentioned.

Chapter 3 starts into the formal results of the thesis. It presents the syntax and semantics of the logic, called **Lp**, used to represent the statistical knowledge. Examples are given of the types of knowledge expressible in this logic.

Chapter 4 presents the deductive proof theory for **Lp**. The proof theory is shown to be both sound and complete, and examples are given of the types of reasoning possible.

Chapter 5 develops the inductive mechanism of belief formation, giving some formal justifications for the mechanism. Examples are presented of how the two pieces, knowledge and deduction in **Lp** and belief formation, work together to solve various problems.

Chapter 6 presents an application of the formalism to a treatment of multiple inheritance with exceptions. A graph based inheritance reasoner is presented, which in some ways is more general than any previous system.

Finally, chapter 7 sums up what has been accomplished and presents some suggestions for future research.

Chapter 2

Introduction to the Formalism

There are two parts to the formalism developed in this thesis. The first is a probabilistic logic, called **Lp**, which is used to represent and reason with a wide range of statistical knowledge. The second is an inductive mechanism of inference which uses this knowledge to generate degrees of belief in a wide class of assertions.

2.1 The logic Lp

2.1.1 Types of Statistical Knowledge

One of the criticisms of the use of probabilities in AI was stated in an influential article by McCarthy and Hayes [48], in which they observed:

> The information necessary to assign numerical probabilities is not ordinarily available. Therefore, a formalism that required numerical probabilities would be epistemologically inadequate.

This has been an on-going and valid criticism of the use of probabilities for general knowledge representation.

This is also the reason why the area in which probabilities have had their major impact has been in specialized expert systems. In such domains numbers (of some degree of accuracy) are sometimes available, and are obtained by interviewing domain experts. The development of methods for structuring probabilities into causal networks (Pearl [55]) has further increased the popularity of using probabilities in expert systems, especially for medical diagnosis.[1] These methods have two main advantages. First, they reduce the quantity of probabilistic information required, without positing overly restrictive assumptions of independence. Second, they reduce the sensitivity of the inferences generated to errors in the original numbers.

However, the impact of probabilities on general knowledge representation remains limited. These approaches still require a significant amount of numerical data, which makes them unsuitable for general knowledge. Furthermore, all of this work has been based on propositional languages, and such languages are inadequate for general knowledge representation.

---

[1] The resurgence of probabilities in expert systems can be traced in a series of articles, starting with Heckerman [26] who reinterpreted MYCIN's mechanism of certainty factors in terms of probabilities, Horvitz et al. [31] pointed out the importance of Cox's normative results, and most recently Heckerman et al. [27] and Schachter et al. [71] discuss the advantages of causal nets.

This work attempts to meet the objection of McCarthy and Hayes by developing a logic which is capable of expressing a wide range of non-numerical probabilistic knowledge; furthermore, if numbers are available they too can be represented. In order to attain this goal the following typology of probabilistic knowledge was constructed:

Relative: Probabilistic knowledge may be strictly comparative. For example, while most would agree that it is more likely that a politician was trained as a lawyer than as an engineer, few would be able to assign values to these probabilities.

Interval: Even when we can give numeric values to the probabilities in question these values may only be in the form of intervals; e.g., the probability of a politician being a lawyer may be in the range 0.7–0.9.

Functional: Probabilistic knowledge can also be in the form of functional information. For example, it would seem that the weight of a bird is a factor in its ability to fly. It is clear that given the weight of a bird we cannot deduce its ability to fly, nor its inability to fly. This knowledge could, however, be expressed as "The probability that a bird can fly is a (decreasing) function of its weight."

Conditional: In a general knowledge base we may have totally unrelated sets of probabilistic knowledge, e.g., "Most dogs bark" along with "Most old cars need repairs." Conditional probabilities can be used to represent the fact that these pieces of information are independent.[2] For example, if we construe 'Most' as meaning more than 50% these statements could be represented as the **Lp** sentences $[Bark(x)|Dog(x)]_x > .5$ and $[Needs\_Repairs(x)|Old\_Car(x)]_x > .5$, where the square brackets are used to indicate probability and x can be considered to be a random member of the predicates' denotations. In the first sentence, for example, no assertion is being made about the probability of a randomly selected object, x, barking unless x is known to be a dog.

Independence: Knowledge of independence is also a type of probabilistic knowledge which people possess. For example, most doctors would agree that the colour of their patients' shoes has no influence on their illnesses. Work by Pearl and his associates has demonstrated the importance of this kind of knowledge ([56], [58], [59]).

The logic developed in this work, **Lp**, deals with all of these considerations. It allows for the expression of all of these different types of knowledge.

---

[2] Hempel [28, page 136] makes a cogent argument that *all* probabilities are in fact conditional probabilities. Indeed, in the logic constructed unconditional probabilities make very little sense, except perhaps, in very circumscribed domains.

Along with all of this probabilistic knowledge it is also clear that some knowledge is certain, and in some situations both certain and probabilistic knowledge must be used to make the desired inferences. For example, if we have the certain knowledge that kiwis are cultivated fruits and also the probabilistic knowledge that most cultivated fruit are edible, we have to use both pieces of information to conclude that kiwis may be edible. A formalism is required which can represent both certain (logical) and uncertain (probabilistic) knowledge before such types of reasoning can be mechanized. **Lp** is an extension of first order logic, so logical knowledge of this nature can also be expressed.

2.1.2 The Field of Numbers

A key, and rather simple innovation which contributed to the expressiveness of the logic was to make it two sorted, by including a totally ordered field of numbers in the semantic model. One sort of entity in the logic is a set of objects, $\mathcal{O}$, and the other sort is a field of numbers, $\mathcal{F}$. The intention is that the set of objects consists of the things of interest (e.g., cars, people, kinds of cars), while the field of numbers consists of ordinary real numbers.[3]

Since the numeric values of the probability terms are part of the object language, it becomes possible to express relationships between them without specifying the actual numbers. For example, it is possible to express the statement "It is more likely that a politician is a lawyer than an engineer" in **Lp** as

---

[3] Using real numbers was the intention, however, technicalities forced the use of an abstract totally ordered field instead (see section 2.1.4).

$$[Lawyer(x)|Politician(x)]_x > [Engineer(x)|Politician(x)]_x.$$

No commitment is made to a specific value for either of the probability terms mentioned, i.e., no value is asserted for either $[Lawyer(x)|Politician(x)]_x$ or $[Engineer(x)|Politician(x)]_x$. Furthermore, if symbols representing some subset of the reals are specified, it becomes possible to express intervals. For example, if the symbols '0.7' and '0.9' are used (and given their normal arithmetic meaning) statements like "The probability that a politician is a lawyer is between 0.7 and 0.9" can be represented by the **Lp** sentence:

$$[Lawyer(x)|Politician(x)]_x \in (0.7, 0.9).$$

Once a field of numbers was added to the logic it became possible to include 'measuring' functions in the logic. These measuring functions map from the set of objects to the field of numbers. Using the measuring functions it is possible to express a statement like "Jack's weight is 80 kilograms." This can be expressed with the **Lp** sentence $Weight\_in\_Kgs(Jack) = 80$, where $Weight\_in\_Kgs$ is defined to be a measuring function and $Jack$ and '80' are constants (object and field constants respectively). The objective of expressing functional probabilistic knowledge was attained by allowing sentences to be constructed recursively from these types of symbols. For example, the statement "Heavier birds are less likely to be able to fly" can be expressed in **Lp** using a measuring function symbol.

2.1.3 Probabilities over the Domain of Discourse

It was the desire to represent logical information which played the key role in the evolution of this work. In a general knowledge base it would be necessary to represent both general knowledge like "All men are mortal" and specific knowledge like "*Socrates* is a man." Propositional logics are inadequate for this purpose. These statements represent distinct assertions, and thus must be encoded as distinct propositional symbols. There is no inherent relationship between distinct symbols in a propositional logic. However, there is clearly a relationship between these different assertions, a relationship which propositional logics cannot express nor reason with. It is necessary to use first order logic (i.e., predicates and quantification) to capture the semantics of such relationships. For example, in first order logic it is possible to use the relationship between the assertions "All men are mortal" and "*Socrates* is a man" to infer "*Socrates* is mortal."

Hence, in order to represent logical information, methods of mixing probabilities with first order logic were examined. Most work in this area has examined methods for attaching probabilities to the sentences of a logical language. The work of Nilsson [53] and Bundy [6] are examples of this approach in AI. They posit a probability distribution over a collection of possible worlds. Each possible world in this collection gives a complete specification of the truth values of the atomic sentences of the logic, and the collection consists of all such distinct truth value specifications. Each sentence of the logic is assigned a probability equal to the measure of the set of possible worlds in which that sentence is true. This approach is, however, incapable of expressing statistical generalizations, e.g., the statement "More than 50% of all dogs can bark."[4]

This statement makes a claim about dogs in general, hence, it cannot be expressed by assigning probabilities to statements about particular dogs. It would seem that some sort of variable is required. The only type of sentences in first order logic which have variables are those which contain a universal quantifier. However, this statement cannot be expressed by assigning a probability to the universal sentence $\forall x Dog(x) \rightarrow Bark(x)$. If the knowledge base contains a *single* instance of an individual dog which cannot bark, i.e., an individual dog with probability zero of barking, the universal sentence will be false in all possible worlds; thus, the probability of the universal sentence will be zero, even if every other dog in the knowledge base is known to bark.

---

[4] To be more precise, it is incapable of representing such statements without using devices like set theory to build up enough mathematics to express statements of proportion.

Recently Fagin et al. [20] have developed a logic for reasoning about probabilities. They give, unlike Nilsson, axioms for reasoning about the probabilities (Nilsson only provided some approximation techniques), and furthermore they prove their axiomatization to be complete. However, their logic is also a method for assigning probabilities to sentences using a possible worlds approach, so it too suffers from the same difficulty when it comes to representing statistical statements.

There has also been a considerable amount of work in philosophy concerned with probability logics (*e.g.*, Carnap [9], Gaifman [22], Field [21], van Fraassen [79], LeBlanc [41], Morgan [50]). This work also has been concerned with attaching probabilities to the sentences of a logic. The possible worlds approach used in AI is equivalent to the approach taken by these philosophers. In particular, they posit a probability distribution over the Lindenbaum-Tarski algebra formed by equivalence classes of sentences in the logic. These equivalence classes are defined by the relation of provable equivalence.[5] The bases for the probability distribution are sentences which are long conjunctions and which fix the truth value of all other sentences. Corresponding to each such long conjunction is a possible world. In fact, each possible world can be identified with a long conjunction which specifies the truth values in that possible world. Hence, a probability distribution over these base sentences is equivalent to a probability distribution over the set of possible worlds. When the logic is first order logic, universally quantified sentences are assigned a probability which is equal to the infimum of the probabilities of all instantiations of that universal. This is called the substitutional interpretation (see LeBlanc [41]). (Technical details differ from author to author.) In fact, this is the only reasonable interpretation if one also wishes to preserve the normal meaning of universal sentences. The substitutional interpretation has the property that one false instantiation (i.e., an instantiation with probability zero) forces the probability of the universal to be zero (cf. the possible worlds approach described above).

---

[5] That is, if $\alpha \rightarrow \beta$ and also $\beta \rightarrow \alpha$, then α and β are in the same equivalence class.

The difficulty with the method of attaching probabilities to the sentences of a first order logic is that the only kind of variable available in a first order logic are universally quantified variables. The device of attaching probabilities to sentences does not change the essential character of such variables. The statement "More than 50% of all dogs can bark" can also be interpreted as saying that a randomly selected dog has a greater than 50% chance of being able to bark. Universally quantified variables are not random variables. The novel feature of the logic developed in this work, **Lp**, is that it has random variables as well as universally quantified variables.

Lp does not have a probability distribution over the sentences of a logical language. In **Lp** the probability distribution is, instead, over the domain of discourse. This is an explicitly empirical interpretation of the probabilities, whereas, the possible worlds approach can be viewed as being a subjective approach to probabilities. The logic is also an extension of ordinary first order logic. In the logic statistical knowledge is expressed through probability terms which contain open formulas (i.e., formulas with free variables). For example, the statement "More than 50% of all dogs bark" can be expressed with the **Lp** sentence "$[Bark(x)|Dog(x)]_x > 0.5$". This sentence is formed from the '$>$' predicate symbol, the constant '0.5', and a probability term which contains two open formulas, $Bark(x)$ and $Dog(x)$. The square brackets are used to form probability terms. These terms are formed by binding some of the free variables of the open formula. In this case the free variable x is bound by the probability term. The variable x used in this manner can be interpreted as being a random variable. Intuitively, the probability term represents the probability that a randomly selected dog, x, will be able to bark. Equivalently, it can be viewed as representing the proportion of objects, x, which bark among those which are dogs. These probability terms have a completely different semantics from the semantics of universal sentences, and can be used to express a wide variety of statistical knowledge.

In **Lp**, however, closed formulas can only have probability one or zero. That is, in **Lp** a closed formula like "$Bark(Fido)$" is either true or false; no intermediate value is possible.

So, it can be seen that the subjective and empirical approaches to probability are in a sense two parts of a complete picture. The possible worlds approach, which expresses a subjective probability, can assign a probability to a closed formula, but is incapable of representing empirical probabilities, which take the form of statistical statements. **Lp**, on the other hand, is capable of expressing these statistical statements, but is incapable of representing a subjective probability assignment to a closed formula.

The generation of subjective probabilities, which can be viewed as being degrees of belief, is the task of the second part of the formalism, belief formation. Belief formation is an inductive inference mechanism which can use the statistical knowledge expressed in **Lp** to generate degrees of belief in a wide class of closed formulas. These degrees of belief are not exactly like subjective probabilities. As mentioned in chapter 1, subjective probabilities need have no relationship with reality. The degrees of belief generated in this formalism, however, are based on objective knowledge of the world, i.e., on the statistical knowledge expressed in **Lp**.

2.1.4 Non-Standard Probabilities

Mathematically standard probabilities have two features beyond the calculus which defines their behaviour. First, they are *real valued* measures, and second they are *sigma-additive*. Sigma-additive probabilities are subject to the constraint that the probability of any countably infinite collection of mutually disjoint sets is equal to the limit of the sum of the individual probabilities of those sets.

The probabilities used in this work are non-standard in both respects. First, their range of values is only required to be a totally ordered field. Second, they are finitely additive but not necessarily sigma-additive. These non-standard features are a result of pragmatic considerations.

One of the constraints under which this work was developed was the desire to keep the proof theory of the logic relatively simple, with the eventual aim of mechanization in mind. This precluded any constructions like infinite rules of inference. Without an infinite rule of inference it is impossible to have a complete proof theory which guarantees sigma-additivity.[6] A trivial way out of this problem is to restrict the domain of discourse to be finite. For finite domains finite additivity trivially corresponds to sigma additivity. This is essentially the route taken by Fagin et al. [20], they restrict their attention to propositional languages where it is impossible to refer to an infinite set (since sentences, being finite in length, can only refer to a finite collection of atomic symbols).

However, the major aim here is an expressive logic for AI, not adherence to standard mathematical practice. There are many interesting concepts, particularly in statistics, which require the notion of at least countably infinite domains, e.g., repeatable trials. The difficulties which arise from probabilities which are not sigma-additive are circumvented, to some extent, through the use a characterization of such probability functions due to Schervish et al. [72] (see chapter 3).

---

[6] Keisler [33] has shown that finite logics with sigma-additive probability distributions over the domain of discourse are not compact. That is, such logics may have an infinite set of sentences which is inconsistent even though every finite subset is consistent.

The other non-standard property is the fact that the probabilities used here are field valued not real valued. Fagin et al. [20] are able to axiomatize real valued probabilities by using the theory of real closed fields. Tarski [76] has shown that this theory is complete for the reals. That is, any sentence in the theory of real closed fields is provable if and only if it is true of the reals. Unfortunately the theory is very expressive—it only allows the constants 0, 1, and −1, and no functions other than addition and multiplication. This means that the 'measuring' functions used here, which play a key role in extending the expressiveness of **Lp**, would not be allowed. Also it would not be possible to make statements which assert that probabilities are functions of other values. So, for example, one could not assert that certain quantities are normally distributed. Hence, the desire for expressiveness again mandated a sacrifice of standard mathematics.

Non-real valued probabilities are much easer to deal with than are non-sigma-additive probabilities. The reals are an example of a totally ordered field; thus anything provable of field valued probabilities will also be true of real valued ones. It is also known that the rationals can be embedded in every totally ordered field, which means that the probabilities can take on any rational values (between 0 and 1) that they wish. This means that *field valued probabilities are sufficient for all practical purposes* since computers are only capable of dealing with rational numbers (and only a finite set of them).

It will be shown later that many interesting statements true of real valued probabilities are also provable of these field valued probabilities. In fact, existent work in AI has only used simple properties of standard probabilities, properties which are also provable of the non-standard probabilities used here. The advantage of using non-standard probabilities are that they allow a very expressive logic with a complete proof theory which is very similar to ordinary first order proof theory.

2.1.5 Related Work

As mentioned in the previous section, previous work on probability logics in artificial intelligence has investigated the attachment of probabilities to sentences; thus, is not directly comparable with this work. It should be noted, however, that **Lp**, along with the mechanism of belief formation, can duplicate much of the reasoning performed by these logics. Furthermore, **Lp** can express a considerably greater variety of probabilistic information than either Nilsson's probability logic or the extensions to Nilsson's logic presented by Grosof in [24].

The work of Grosof [25] represents an alternative approach to getting non-monotonic behaviour out of probabilities. Starting from sentences (closed formulas) in an ordinary first order language he constructs a new language in which these sentences are terms. The sentences in the new language are assertions which assign probabilities to the sentences in the original first order language. He then adds non-monotonic features to his new language. These non-monotonic features are formalized using Lifschitz's pointwise circumscription [44], which requires the use of higher order languages. Besides the complexity of this approach, it is also the case that the mechanism of assigning probabilities to sentences of first order logic is incapable of expressing statistical knowledge. Hence, it is not clear how this scheme can be used in a knowledge-based reasoning system which uses statistical knowledge. In **Lp** there is no need for higher order languages to attain non-monotonicity; belief formation yields non-monotonicity using just the classical laws of conditional probabilities.

The work which is most closely related to the logic **Lp** is the work of the mathematician Keisler [33]. His work lays the foundations for probability logics where the probability distribution is defined over the domain of discourse, as is the case for **Lp**. The aim of his work is, however, to develop a logic for expressing mathematical notions where uncountable domains of discourse are common. Keisler has shown that when the domain is uncountable a logic cannot be coherent if it possesses both probabilities over the domain of discourse and universal quantification.[7] Uncountable domains are not, however, of paramount importance in AI, where we are primarily concerned with statements about the 'ordinary objects' of human experience. Thus, restricting the class of admissible models to be at most countably infinite in cardinality has enabled the development of a logic **Lp** which for AI is far more expressive than Keisler's logic. The logic allows both universal quantification as well as a probability distribution over the domain of discourse. In fact, **Lp** is an *extension* of ordinary first order logic; thus, it can represent all statements expressible in first order logic, as well as probabilistic knowledge. Furthermore, Keisler has no need for an inductive mechanism, and does not address this problem.

There has also been some work in the philosophy of language which is similar to the logic **Lp**. Åqvist et al. [1] give a semantic analysis of adverbs of frequency (e.g., always, sometimes, often). Their semantic model is essentially a first order logic with a probability function over the domain of discourse. They, however, restrict themselves to finite models and a less expressive logic. Furthermore, they do not address the problem of induction. Hence, their formalism can represent sentences which contain adverbs of frequency, but it cannot reason inductively with these sentences.

---

[7] This limitation arises from the fact that the projection sets generated through universal quantification may not, in general, be in the domain of any probability function, i.e., they may be nonmeasurable sets.

Another important difference between the logic **Lp** and both Keisler's logic and the formalism of Åqvist et al. is the field of numbers in the semantic model. In order to refer to the values of the probabilities, Keisler uses a device he calls probability quantifiers (P-quantifiers). This device is similar to the so called J-operators which are standard in many valued logics (see Rosser and Turquette [66]). The intent of this device is to give access in the object language (syntax) to the semantic probabilities. For example, one can write the sentence $(Px = 0.5)\theta$ to indicate that the proportion of objects for which θ is true (when the variable x in θ is interpreted as that object) is 50%. These P-Quantifiers become part of the fixed logical symbols of the language. With these P-Quantifiers, however, the numerical values of the probabilities remain outside the main part of the logic. That is, the numbers (like 0.5) which appear inside the P-Quantifiers cannot be referred to outside of the P-Quantifiers. As a result arithmetic relationships between probabilities cannot be expressed. For example, it is not possible to express the previous statement: "It is more likely that a politician is a lawyer than an engineer" using P-Quantifiers. This statement cannot be expressed without a commitment to the values of the probabilities of both cases. That is, we could say something like $(Px = .8)(Lawyer(x)|Politician(x))$ and $(Px = .4)(Engineer(x)|Politician(x))$

but not something like

$$(Px)(Lawyer(x)|Politician(x)) > (Px)(Engineer(x)|Politician(x)),$$

as this is not a valid form of the P-Quantifier. Åqvist et al. [1] are not concerned with exact numbers, and rely on predicates like 'usually', 'sometimes', etc., to express the numerical import of their sentences; no arithmetic relationships between these predicates can be represented.

2.2 Belief Formation

Intuitively it is clear that a sentence like "*Bark*(*Fido*)" is objectively either true or false; however, when the actual truth value is unknown it may be necessary to make a reasonable guess. This is the purpose of the inductive mechanism of belief formation. This mechanism generates a 'reasonable' guess which is based on the statistical knowledge expressed in **Lp**. This guess is in the form of an assignment of a probabilistic degree of belief to the closed formula. These induced degrees of belief are not to be confused with statements in **Lp**. **Lp** is incapable of expressing such degrees of belief. It is, however, capable of expressing base statistical information from which these degrees of belief can be generated.

2.2.1 The Inductive Assumption of Randomization

The mechanism is founded on a very simple inductive assumption which has a long history. The process is similar to the way in which we make sense of statements like "The probability that a coin will show heads when flipped is 0.5". For a particular instance of flipping a coin it is necessarily the case that the coin will show either heads or tails, i.e., the truth value of "This flip will show heads" (abbreviated as *Show_heads*) will be either zero or one. When we state that the probability of *Show_heads* is 0.5 we are implicitly randomizing this particular coin flip. In other words, we know that 50% of the instances of flipping various coins yield heads (assuming that there are as many coins biased to heads as to tails), and, since we do not have any information that distinguishes this particular coin flip from any other coin flip, it is reasonable to believe *Show_heads* to degree 0.5, by assuming that this is a randomly instance of a coin flip. Similarly for the example of "*Bark*(*Fido*)," if it is known that (say) 90% of all dogs bark and all that is known about *Fido* is that he is a dog, then the inductive assumption, that *Fido* is a randomly selected dog, would impart a degree of belief of 0.9 to the (closed) formula "*Bark*(*Fido*)".

These levels of belief are justified in the long term. For example, consider the assignment of 0.5 as a degree of belief in the assertion *Show_Heads*. If one was to make bets based on this degree of belief, one would never accept odds lower than 1:1 that a coin flip will show heads, unless one had reason to believe that the particular coin was biased. These odds are acceptable, since in the long run, over a sequence of bets, one would break even. If the level of belief was different from 0.5, while at the same time the long term frequency of heads was 50%, then one could accept odds, either for heads or for tails, which would lead to eventual ruin.

Inductive assumptions of randomization have appeared before in the philosophy of science literature, at least as early as Reichenbach (1949 [64]) and more recently in work by Kyburg [37,40]. Similar inductive assumptions have also appeared *implicitly* in most of the expert systems which deal with uncertainty. For example, in the MYCIN system most of the rules which have certainty factors are in the form "The certainty of infection D given symptoms A, B and C is x." These certainty factors are synopses of an expert's experience with a population of patients. When diagnosis is performed on a particular patient it is assumed that these certainty factors are applicable to that patient. Here an implicit randomization identical to an inductive assumption is taking place; the particular patient is assumed to be a random member of the population of patients.

2.2.2 Conflicting Knowledge

The inductive assumption must deal with situations where the knowledge available is conflicting. The previous example of Fido barking can be used to illustrate the point. The inductive assumption generates a single degree of belief for the sentence "$Bark(Fido)$" only when all that is known about Fido is that he is a dog (i.e., all that is deducible about Fido from the knowledge base using **Lp** deduction). This situation is rare; usually much more is known about named individuals. For example, the sentences "$Dingo(Fido)$" or "$Brown(Fido)$" may also be in the knowledge base. In general, the degree of belief in "$Bark(Fido)$" induced from the knowledge that Fido is a dog will be completely different from the degree of belief induced from other knowledge about Fido, say for example, the knowledge that Fido is a dingo. That is, considering Fido to be a randomly selected dog yields a different degree of belief than when Fido is considered to be a randomly selected dingo (dingos don't bark). Thus, the knowledge base can generate a range of different degrees of belief for any sentence, dependent on what knowledge is used in the inductive step of randomization.

For example, if the knowledge base contains the assertions "Most Republicans are not pacifists", "Most Quakers are pacifists", "Nixon is a Republican", and "Nixon is a Quaker", where 'Most' is interpreted as meaning greater that 50%, belief formation can generate two degrees of belief in the assertion "Nixon is a pacifist". The degree of belief in this assertion given the knowledge that Nixon is a Republican will be more that 0.5, while the knowledge that Nixon is a Quaker will generate a degree of belief of less than 0.5. In some situations there will be no way to choose between these differently founded degrees of belief; however, in many situations there is a simple preference criterion which can be applied.

This preference criterion is based on the simple intuition that the more knowledge that is used to generate the degree of belief the better is that degree of belief. More knowledge has a simple interpretation in **Lp**; the sentence α represents more knowledge than β if $\alpha \rightarrow \beta$ is deducible from the knowledge base.[8] This allows us to use deduction in **Lp** in two different ways when forming beliefs. First, deduction can be used to generate degrees of belief in sentences for which there is no explicit statistical knowledge, through the deduction of new statistical knowledge. Second, deduction can sometimes be used to decide between competing degrees of belief generated from different knowledge. The assumption that subclasses should override superclasses, which is normally used in inheritance hierarchies (Touretzky [77]), is generalized by this preference criterion. The generalization arises from the fact that the criterion is not restricted to a limited notion of classes, but instead, is applicable to any 'class' defined by arbitrarily complex formulas in **Lp**. Furthermore, the preference criterion can be given a formal justification based on the semantics of **Lp**.

---

[8] In general it is undecidable in first order logic (and thus in **Lp**) as to whether or not $\alpha \rightarrow \beta$ is deducible from the knowledge base. One can, however, always make the conservative assumption that $\alpha \rightarrow \beta$ is *not* deducible if a deduction is not found before some resource limit is exceeded. By assuming that no deduction exists one is forced to consider both degrees of belief.

2.2.3 Kyburg's Work

Kyburg [37] has developed a formalism which is very similar in its philosophy to the system constructed here. His work is an attempt to give a definitive treatment of the logic of statistical inference, removing many of the philosophical and logical shortcomings of classical statistics. In contrast, **Lp** and its associated mechanism of belief formation were arrived at through an investigation of the mathematical consequences of mixing probabilities with first order logic, rather than through an inquiry into the philosophical foundations of statistical inference. Kyburg constructs a system in which probabilities, or degrees of belief as they are called here, are assigned to particular assertions based on some underlying statistical knowledge. This approach is similar to the system of belief formation presented here. The chief differences are that, first, he uses a complex meta-language (which includes all of Zermelo-Fraenkel set theory) to express the statistical knowledge, whereas here, the statistical knowledge is expressed in the object language **Lp** (clearly more suitable for our eventual aim of automation), and second, he develops a complex set of preference criteria (in [40]) for choosing a single most preferred degree of belief, whereas here, there is only the single preference criterion of implication, which yields only a partial order between competing degrees of belief. Furthermore, this preference criterion uses the proof theory of the object language to generate its preferences rather than expressions in a meta-language, as used by Kyburg.

2.3 A Simple Rational Agent

The degrees of belief can be used by a rational agent to guide action in situations where the actual truth value of an assertion is unknown. A simple model of a rational agent can be developed which demonstrates where the logic **Lp** and the mechanism of belief formation fit in. There are three components to this simple model, a goal directed planner, a knowledge base expressed in the logic **Lp**, and the mechanism of belief formation.

The planning component interacts with the environment performing actions intended to attain its goals and receiving new information which is added to the knowledge base. To be rational the agent must use its knowledge of the environment. If certain things are true in the environment then the agent must perform certain actions. For example, if the agent wants to cross a road and there is a car coming, the agent must wait until the car has passed in order to satisfy goals of preservation. There are many situations, however, where knowledge is incomplete; for example, there may be a blind corner which makes it impossible for the agent to detect oncoming cars. In this situation the agent may use its statistical knowledge (also part of the same **Lp** knowledge base) along with belief formation to assign a reasonable degree of belief in the assertion that a car is coming. The statistical knowledge may be knowledge of the frequency of traffic around that particular corner, knowledge of the frequency of traffic at that time of day, etc. The degree of belief generated through belief formation can then be used by the planning component to make a rational decision, a decision which considers, for example, the risk of crossing the road at that point and the costs of moving to a safer spot.

It will be shown that if the truth value of the sentence is known, i.e., the sentence or its negation is deducible from the knowledge base, then the degree of the most preferred belief assignable by the mechanism of belief formation will be representative of that truth value, i.e., it will be 0 or 1 as the truth value is false or true respectively. So, there is no need to distinguish between situations where knowledge is certain and where it is not. The agent always uses its beliefs to guide its actions, beliefs which are in turn generated by reasoning with its knowledge.

Since **Lp** subsumes first order logic, it is undecidable. Hence, this "most preferred belief" will not always be generatable, since it will never be known when to stop looking for a deduction of the sentence. As most rational agents operate under time constraints, a reasonable thing to do is to use the beliefs which have been generated up to that time limit.

To insure that such a "real time" agent behaves rationally it would be necessary for it to organize its knowledge base in an efficient manner. That is, facts which are most relevant to the formation of crucial beliefs (i.e., beliefs which guide important actions) must be quickly deducible. Specialized inference structures like taxonomic hierarchies, or causal networks, play an important role in this regard. This point will be further examined in chapter 6.

As the agent interacts with its environment it learns new facts, which are added to the knowledge base. These new facts can change the agent's beliefs. That is, the degrees of belief generated through the inductive mechanism change as new knowledge is added to the knowledge base; they exhibit non-monotonic behaviour. For example, the agent may have a degree of belief > 0.5 in the assertion that *Fido* can bark, based on knowledge about *Fido*. If later the agent actually hears *Fido* barking, i.e., the assertion is added to the knowledge base, then the belief formation mechanism will generate a new preferred degree of belief, equal to one, in the assertion.

Chapter 3

The Logic Lp

This chapter presents the syntax and semantics of the logic **Lp**. The formalization of **Lp** follows the standard steps used in the development of ordinary first-order logic (see for example Bell [3]). First, the set of allowed symbols is defined. Then rules are given which specify the strings of symbols which are the well-formed formulas. This defines the syntax of **Lp**. Next, the semantics of **Lp** are given, by first defining the set of admissible models (**Lp**-Structures), then a correspondence between truth in the models and the well formed formulas. In the next chapter a deductive proof theory is presented which provides a correspondence between truth in the model and a syntactic manipulation of the formulas. The deductive proof theory is shown to be both sound and complete.

The letters n and m are used as variables which refer to natural numbers.

3.1 Symbols

The following are the symbols of the language **Lp**. The total number of symbols is denumerable.

a) A set of constant symbols $(a, b, c, \ldots)$.

b) A set of function symbols $(f, g, h, \ldots)$.

c) A set of predicate symbols $(P, Q, R, \ldots)$.

d) A set of variables $(x, y, z, \ldots)$.

For each of (a)-(d) there are two types of symbols, object symbols and field symbols. The field symbols will be written in a **bold** font when there is a danger of confusion

e) A set of measuring function symbols ($Weight$, $Size$, μ, ν, ρ, $\ldots$).

Also included are the distinguished symbols of **Lp**:

a) The binary object predicate symbol $=$.

b) The field constant symbols 1 and 0, the field binary predicate symbols $\geq$ and $=$[1], and the field binary function symbols $+$, $\times$, $-$, and $\div$.

c) The connectives $\wedge$ and $\neg$.

---

[1] Note, '=' is used both as a field and as an object equality symbol. This should not, however, cause any confusion.

d) The quantifier $\forall$.

e) The probability term former $[\circ]$.

3.2 Formulas

The formulas of **Lp** are strings of **Lp** symbols formed by the following recursive rules. The formulas constructed by these rules are the only formulas of **Lp**.

T0) A single object variable or constant is an *o-term*; a single field variable or constant is an *f-term*.

T1) If f is an n-ary object function symbol and $t_1, \ldots, t_n$ are o-terms, then $ft_1 \ldots t_n$ is an *o-term*. If $\mathbf{f}$ is an n-ary field function symbol and $\mathbf{t}_1, \ldots, \mathbf{t}_n$ are f-terms then, $\mathbf{ft}_1 \ldots \mathbf{t}_n$ is an *f-term*. If ν is an n-ary measuring function symbol and $t_1, \ldots, t_n$ are o-terms, then $\nu t_1 \ldots t_n$ is an *f-term*.

F1) If P is an n-ary object predicate symbol and $t_1, \ldots, t_n$ are o-terms, then $Pt_1 \ldots t_n$ is a *formula*.

F2) If $\mathbf{P}$ is an n-ary field predicate symbol and $\mathbf{t}_1, \ldots, \mathbf{t}_n$ are f-terms, then $\mathbf{Pt}_1 \ldots \mathbf{t}_n$ is a *formula*.

F3) If α is a formula then so is $\neg\alpha$.

F4) If α and β are formulas then so is $\alpha \wedge \beta$.

F5) If α is a formula and x is a variable (of either type), then $\forall x \alpha$ is a *formula.*

T2) If α is a formula and $\vec{x}$ is a vector of n object variables $\langle x_1, \ldots, x_n \rangle$, then $[\alpha]_{\vec{x}}$ is an *f-term.*

This definition of formulas is different from the standard first order definition; the last rule of formation allows terms to be constructed from formulas.

The connectives $\vee$ and $\rightarrow$, and the quantifier $\exists$ are defined in the standard manner from the given primitives. The predicate symbols $=$ and $\geq$ as well as the function symbols $+$, $\times$, $-$, and $\div$, are written in the more readable infix form. Furthermore, standard conventions of scope and precedence are used to limit the use of parentheses. It is also convenient to introduce the following abbreviations to express inequalities between field terms.

Definition 3.2.1

a) $x \leq y =_{df} y \geq x$ *b)* $x \in (y, z) =_{df} y \leq x \wedge x \leq z$

c) $x < y =_{df} \neg(x \geq y)$ *d)* $x > y =_{df} \neg(y \geq x)$

Conditional probabilities are represented in **Lp** through the following abbreviation.

Definition 3.2.2 $[\alpha|\beta]_{\vec{x}} =_{df} [\alpha \wedge \beta]_{\vec{x}} \div [\beta]_{\vec{x}}$.

Note, in this definition there is no mention of what happens if $[\beta]_{\vec{x}} = 0$. The reason is that this is a syntactic abbreviation, and there is no way of determining *syntactically* if a term such as $[\beta]_{\vec{x}}$ is equal to zero. This will be determined by the underlying semantic model; i.e., in some models this term will be equal to zero in others it will not. In fact, it can be seen that this problem occurs with any use of the division function. That is, it is impossible to detect division by zero syntactically. This problem is dealt with pragmatically—the result of division by zero is left undetermined. In any model the division function will be total, that is, the division function will give a particular result when the divisor is zero, but this result can be anything and can vary from model to model. Division by non-zero numbers will, however, behave in the expected manner in all models. When necessary it is always possible to *guard* against division by zero syntactically, by including an explicit conditional. For example, one could write

$$[\beta]_{\vec{x}} \neq 0 \rightarrow [\alpha|\beta]_{\vec{x}} > [\delta|\beta]_{\vec{x}}.$$

Here the initial implication acts as a guard against division by zero.

3.3 Semantic Model

Definition 3.3.1 (The Model) *An* Lp-*Structure is defined to be the tuple*

$$\mathcal{M} = \left\langle (\mathcal{O}, R_{\mathcal{O}}, F_{\mathcal{O}}), (\mathcal{F}, R_{\mathcal{F}}, F_{\mathcal{F}}), \Psi, \{\Pi_n, \mu_n \mid n = 1, 2 \ldots\} \right\rangle$$

Where:

a) $(\mathcal{O}, R_{\mathcal{O}}, F_{\mathcal{O}})$ represents a ***countable*** set of individual objects $\mathcal{O}$, a set $R_{\mathcal{O}}$ of relations on the objects, including the equality relation, and a set $F_{\mathcal{O}}$ of functions from tuples of objects to objects $(\mathcal{O}^n \mapsto \mathcal{O})$.

b) $(\mathcal{F}, R_{\mathcal{F}}, F_{\mathcal{F}})$ represents a totally ordered field of numbers $\mathcal{F}$, a set $R_{\mathcal{F}}$ of relations on the numbers, including the equality relation and ordering relation greater than or equal, and a set $F_{\mathcal{F}}$ of functions from tuples of numbers to numbers including the field operations addition, multiplication, division and negation. $\mathcal{F}$ contains two distinguished elements which are the units of addition and multiplication. In the field of real numbers these units are called zero and one, and the same names will be used to refer to the units of $\mathcal{F}$.

c) Ψ represents a set of measuring functions, functions from $\mathcal{O}^n$ to $\mathcal{F}$.

d) Each Π_n $(n = 1, 2, \ldots)$ is a field of subsets of $\mathcal{O}^n$. This field contains all singleton sets of $\mathcal{O}^n$, i.e., every singleton n-tuple. It also contains all subsets of $\mathcal{O}^n$ defined by the formulas of **Lp** (later the semantic definition of the formulas will give a more precise characterization of these subsets). This field of subsets acts as the domain of the probability function μ_n.

e) $\{\mu_n \mid n = 1, 2, \ldots\}$ is a sequence of probability functions. Each μ_n is a set function whose domain is Π_n, whose range is $\mathcal{F}$, and which satisfies the axioms of a finitely additive probability function (i.e., $\mu_n(A) \geq 0$, $\mu_n(A \cup B) = \mu_n(A) + \mu_n(B)$ if $A \cap B = \emptyset$, and $\mu_n(\mathcal{O}^n) = 1$).

This sequence of probability functions is subject to some further constraints. These constraints ensure that the probability terms behave coherently. The implications of these constraints are discussed in the next section.

1. The sequence of probability functions is a sequence of product measures. That is, for any two sets $A \in \mathcal{O}^n$ and $B \in \mathcal{O}^m$, and their Cartesian product $A \times B \in \mathcal{O}^{n+m}$, if $A \in domain(\mu_n)$ and $B \in domain(\mu_m)$, then

   $$A \times B \in domain(\mu_{n+m}) \quad \text{and} \quad \mu_{n+m}(A \times B) = \mu_n(A) \times \mu_m(B).$$

   The implication of this constraint will be discussed in the next section, but for now it can be noted that this constraint is not a restrictive assumption of independence, like those which have appeared in previous work (see Johnson [32]).

For models where the domain of discourse is *finite* this constraint is sufficient. However, there are many natural notions which involve countably infinite sequences of events or individuals. For example, infinite sequences of trials are often referred to in the study of statistics. In order to include this generality in the logic the probability functions must be sigma-additive, i.e., the measure of the union of a countably infinite collection of disjoint sets must be equal to the limit of the sum of their individual probabilities. Sigma-additivity ensures the probability functions are well behaved in the limit.

As noted in chapter 2 enforcing this constraint presents a difficulty. That is, it cannot be guaranteed that the probability functions are sigma-additive unless infinite rules of inference are allowed. To avoid the complexities arising from infinite rules of inference, weaker constraints are placed on the probability functions. When $\mathcal{O}^n$ is finite these weaker constraints can be deduced directly from the two facts: (a) every singleton set in $\mathcal{O}^n$ is μ_n measurable, and (b) the μ_ns are product measures. When $\mathcal{O}^n$ is countably infinite the additional condition sigma-additivity is needed to derive these constraints.

The weaker constraints have the advantage (over sigma-additivity) of being expressible as axioms in the logic. However, they do admit a larger class of **Lp** models than would sigma-additivity. This situation is similar to the use of an abstract field. That is, the class of **Lp** models includes models in which the field is not the field of real numbers, and similarly it includes models in which the probability functions are not completely sigma additive. The weaker conditions do, however, ensure that the probability terms in **Lp** have properties which are sufficiently similar, for our purposes, to the properties that they would have under the stronger condition of sigma-additivity. There are two additional constraints which are imposed on the probability functions.

The first constraint is that the μ_n measures remain invariant under permutation, even when $\mathcal{O}^n$ becomes infinite. This has the effect that the value of the probability terms are unaffected by the order of the cited variables.

The second constraint can be derived from a characterization of non-sigma-additivity due to Schervish et al. [72]. They have shown that probability functions which are finitely additive but not sigma-additive are characterized by a condition called non-conglomerability.

Let $\pi = \{h_i\}_{i=1}^{\infty}$ be a partition of $\mathcal{O}^n$. The probability measure μ_n is said to be *conglomerable* in π when for every set E in the domain of μ_n for which

$$\frac{\mu_n(E \cap h_i)}{\mu_n(h_i)}$$

is defined for all i, and for all numbers $\mathbf{z}_1$, $\mathbf{z}_2$, if

$$\mathbf{z}_1 \leq \frac{\mu_n(E \cap h_i)}{\mu_n(h_i)} \leq \mathbf{z}_2$$

for all $h_i \in \pi$, then

$$\mathbf{z}_1 \leq \mu_n(E) \leq \mathbf{z}_2.$$

That is, conglomerability asserts that, for each set E, if all the conditional probabilities over a partition π are bounded by two quantities, $\mathbf{z}_1$ and $\mathbf{z}_2$, then the unconditional probability for that event is likewise bounded by the same quantities. The notion of conglomerability is due originally to de Finetti [13].

Schervish et al. have shown that probability functions which are finitely additive but not sigma-additive can be precisely characterized by their failure to be conglomerable over all denumerable partitions. Of course conglomerability over *all* partitions cannot be guaranteed axiomatically, otherwise we would have succeeded in capturing sigma-additivity axiomatically. However, conglomerability over a large class of partitions can be guaranteed through an additional constraint which *can* be expressed axiomatically. This constraint guarantees the conglomerability of conditional probabilities over certain partitions. That is, it guarantees that conditional probabilities will be within certain bounds if they are within those same bounds when subdivided over certain partitions.

To be precise, the two conditions are as follows:

2. Each μ_n is invariant under permutations. That is, for every permutation π of $\{1, \ldots, n\}$ and $S \in domain(\mu_n)$ if

$$\pi S = \{(a_{\pi(1)}, \ldots, a_{\pi(n)}) : (a_1, \ldots, a_n) \in S\},$$

then

$$\pi S \in domain(\mu_n) \quad \text{and} \quad \mu_n(\pi S) = \mu_n(S).$$

3. The probability functions satisfy a conglomerability condition. Let A^{n+m} and B^{n+m} be two sets in $\mathcal{O}^{n+m}$ which are in the domain of μ_{n+m}, and such that the projections of $A^{n+m} \cap B^{n+m}$ and B^{n+m} into $\mathcal{O}^n$ are identical, i.e.,

$$\begin{aligned}(A^{n+m} \cap B^{n+m})^n &= \{\vec{a} | \exists \vec{c} (\langle \vec{a}, \vec{c} \rangle \in A^{n+m} \cap B^{n+m})\} \\ &= \{\vec{b} | \exists \vec{c} (\langle \vec{b}, \vec{c} \rangle \in B^{n+m})\} = B^n.\end{aligned}$$

Furthermore, let the conditional probability of A^{n+m} given B^{n+m} satisfy certain bounds over the partition generated by the first n dimensions: If there exists two numbers $\mathbf{z}_1$ and $\mathbf{z}_2$ such that for every vector $\vec{a}$ in $(A^{n+m} \cap B^{n+m})^n$ $(= B^n)$

$$\mathbf{z}_1 \leq \frac{\mu_m\{\vec{c} | \langle \vec{a}, \vec{c} \rangle \in A^{n+m} \cap B^{n+m}\}}{\mu_m\{\vec{c} | \langle \vec{a}, \vec{c} \rangle \in B^{n+m}\}} \leq \mathbf{z}_2,$$

then,

$$\mathbf{z}_1 \leq \frac{\mu_{n+m}(A^{n+m} \cap B^{n+m})}{\mu_{n+m}(B^{n+m})} \leq \mathbf{z}_2.$$

This condition says that if the conditional probability of A^{n+m} given B^{n+m} is bounded by the numbers $\mathbf{z}_1$ and $\mathbf{z}_2$ over the partition defined by ~~the~~ vectors in the first n dimensions, then the unpartitioned conditional probability also respects the same bounds.

3.4 The Effect of the Coherence Constraints

The sequence of probability functions is constrained to be a sequence of product measures. This insures that distinct variables bound by the probability term formers behave in an independent manner. This is similar to the independence of distinct universally quantified variables in first order logic, e.g., the sentence $\forall x \forall y P(x) \land Q(y)$ can be decomposed into two independent sentences, i.e., $\forall x P(x)$ and $\forall y Q(y)$. Since y and x are distinct variables bound by separate quantifiers, their meanings are independent of each other.

With independence we have, for example, that the probability terms are unaffected by tautologies, e.g., $[P(x) \land (R(y) \lor \neg R(y))]_{\langle x,y \rangle} = [P(x)]_{\langle x \rangle}$.

It should be noted that this constraint on the probability functions does not make any implicit assumptions of independence of the form commonly found in probabilistic inference engines (e.g., the independence assumptions of the Prospector system [17], see Johnson [32]). This constraint affects the values of probability terms with distinct variables, also, complex probability terms, e.g., $[[\alpha]_x = \mathbf{z}]_y$. (This can be seen from axiom (P7), presented in the next chapter, which expresses the constraint.) The constraint does not, however, make any presumptions concerning the independence of formulas which contain the same set of probability variables. That is, in general, $[\alpha \land \beta]_{\mathbf{x}} \neq [\alpha]_{\mathbf{x}} \times [\beta]_{\mathbf{x}}$.

In fact, the probabilistic knowledge that we wish to express in **Lp** normally makes some claim of correlation between the properties possessed by the same object (or tuple of objects). For example, correlation between the properties of being a bird and being able to fly. In this example, the correlation can be expressed by the probability term $[Fly(x)|Bird(x)]_x$, where the same variable appears in both formulas. This probability term expresses the ratio of flying birds among birds. This can be contrasted with the probability term $[Fly(y)|Bird(x)]_{(x,y)}$. In this term the variables are distinct, and its semantic meaning is that we have chosen pairs of objects and are expressing the ratio of the pairs in which the first object is a bird while the second object can fly to the pairs in which the first object is a bird irrespective of the properties of the second object. Since we are referring to different objects, there is no reason for there to be any correlation between their properties.

Correlations between the properties of a particular *set* of objects can be expressed through the use of n-place predicates. For example, the probability term

$$[(Boy(x) \wedge Girl(y)) \vee (Girl(x) \wedge Boy(y)) | Loves(x, y)]_{(x,y)}$$

is not, in general, equal to the product any simpler probability terms.

The second condition, invariance under permutations, ensures, for example, that the order of the variables cited in the probability terms makes no difference, e.g., $[\alpha]_{x,y} = [\alpha]_{y,x}$. Universal quantification also displays this property, e.g., $\forall x \forall y \alpha = \forall y \forall x \alpha$. For finite domains the fact that every singleton set is measurable ensures that all sets of objects are measurable, since the probability function is finitely additive. This along with the product measure constraint ensures that the measures are invariant under permutations. For example,

$$\mu_2\{\langle a, b\rangle\} = \mu_1\{a\} \times \mu_1\{b\} = \mu_2\{\langle b, a\rangle\}.$$

However, this condition must be made explicit for infinite domains.

The third condition is also true in finite domains. For finite domains the partition along the first n dimensions will be a finite one, hence the unpartitioned conditional probability will simply be the weighted average of the probabilities over the partition. The weighted average of a set of numbers all of which are within certain bounds will be within the same bounds. As discussed before, this condition ensures that the probability functions are conglomerable over a large class of partitions in infinite domains.

Another example of the coherence of the probability terms is that they are invariant under variable name changes, e.g., $[P(x)]_x = [P(y)]_y$. This behaviour comes from the manner in which the semantics of the formulas is defined.

3.5 Semantics of Formulas

Meaning is given to the formulas of **Lp** by defining a correspondence between the formulas and the **Lp**-Structure, $\mathcal{M}$, augmented by the truth values $\top$ and $\perp$ (true and false). Such a correspondence is called an interpretation. An interpretation assigns to every object constant symbol an element of $\mathcal{O}$, to every n-ary object function symbol an n-ary element of $F_{\mathcal{O}}$, and to every n-ary object predicate symbol an n-ary element of $R_{\mathcal{O}}$, mapping the distinguished predicate symbol '=' to the equality relation. Similarly, it maps the field constant, function and predicate symbols to elements of $\mathcal{F}$, $F_{\mathcal{F}}$, and $R_{\mathcal{F}}$ respectively, mapping the distinguished symbols 1, 0, $+$, $\times$, $\div$, $-$, $\geq$, and $=$, to the expected constants, operations, and relations in $\mathcal{F}$, $F_{\mathcal{F}}$, and $R_{\mathcal{F}}$. It maps each measuring function symbol to an element of Ψ. Finally, it assigns to each object variable x an element of $\mathcal{O}$ and to each field variable $\mathbf{x}$ an element of $\mathcal{F}$.

These assignments serve as the inductive basis for an interpretation of the formulas. Two interpretations σ and τ are said to **agree** on a given symbol θ if $\theta^\sigma = \theta^\tau$, where θ^σ denotes the interpretation of θ under σ. Also, σ and τ are said to have the same **underlying structure** if they agree on all constant, predicate, and function symbols (of all types). Let $\sigma(x/u)$ denote a new interpretation which is identical to σ except that it assigns the individual u to the variable x (types must match). More generally, let

$\sigma(\vec{x}/\vec{a})$, where $\vec{a} = \langle a_1, \dots, a_n \rangle$ and $\vec{x} = \langle x_1, \dots, x_n \rangle$ are vectors of individuals and variables (of matching type), denote a new interpretation identical to σ except that $(x_i)^{\sigma(\vec{x}/\vec{a})} = a_{i\,(i=1,\dots,n)}$. An interpretation σ is extended to a truth value interpretation of the formulas of **Lp** in the following recursive manner:

T0) If x is a variable or constant (of either type) then x^σ is already defined.

T1) If f is an n-ary function symbol (of either type) and $t_1, \dots, t_n$ are terms of the same type, or if f is an n-ary measuring function symbol and $t_1, \dots, t_n$ are o-terms, then

$$(ft_1 \dots t_n)^\sigma = f^\sigma(t_1^\sigma \dots t_n^\sigma).$$

F1) If P is an n-ary predicate symbol (of either type) and $t_1, \dots, t_n$ are terms of the same type then

$$(Pt_1 \dots t_n)^\sigma = \begin{cases} \top & \text{if } \langle t_1^\sigma, \dots, t_n^\sigma \rangle \in P^\sigma, \\ \bot & \text{otherwise.} \end{cases}$$

F1$^=$) If s and t are terms of the same type then

$$(s = t)^\sigma = \begin{cases} \top & \text{if } s^\sigma = t^\sigma, \\ \bot & \text{otherwise.} \end{cases}$$

F2) For every formula α,

$$(\neg\alpha)^\sigma = \begin{cases} \top & \text{if } \alpha^\sigma = \bot, \\ \bot & \text{otherwise.} \end{cases}$$

F3) For every pair of formulas α and β,

$$(\alpha \wedge \beta)^{\sigma} = \begin{cases} \top & \text{if } \alpha^{\sigma} = \top \text{ and } \beta^{\sigma} = \top, \\ \perp & \text{otherwise.} \end{cases}$$

F4a) For every formula α and object variable x,

$$(\forall x \alpha)^{\sigma} = \begin{cases} \top & \text{if } \alpha^{\sigma(x/a)} = \top \text{ for every } a \in \mathcal{O}, \\ \perp & \text{otherwise.} \end{cases}$$

F4b) For every formula α and field variable $\mathbf{x}$,

$$(\forall \mathbf{x} \alpha)^{\sigma} = \begin{cases} \top & \text{if } \alpha^{\sigma(\mathbf{x}/u)} = \top \text{ for every } u \in \mathcal{F}, \\ \perp & \text{otherwise.} \end{cases}$$

T2) For every formula α the f-term created by the probability term former, $[\circ]_{\vec{x}}$, is given the interpretation,

$$([\alpha]_{\vec{x}})^{\sigma} = \mu_n\{\vec{a} | \alpha^{\sigma(\vec{x}/\vec{a})} = \top\}.$$

Since μ_n is a probability function which maps to the field of numbers $\mathcal{F}$, it is clear that $[\alpha]_{\vec{x}}$ denotes an element of $\mathcal{F}$ under any interpretation σ; thus, it is a valid f-term. As mentioned before, the domain of μ_n is the a field of subsets of $\mathcal{O}^n$ which includes those subsets defined by the formulas of **Lp**. Hence the above set is in the domain of μ_n.

3.6 Examples of Knowledge Representable in Lp

Now we present some examples of knowledge which can be represented in Lp:

Example 3.1 *Notions of typicality.*

"Most birds can fly:"

$$[fly(x)|bird(x)]_x > 0.5,$$

where '> 0.5' is the least presumptive reading of 'Most'.

Example 3.2 *Functional probabilistic relations.*

"Heavier birds are less likely to be able to fly:"

$$\forall \mathbf{y}\Big(Reasonable\_Weight(\mathbf{y}) \rightarrow [fly(x)|bird(x) \wedge weight(x) < \mathbf{y}]_x > [fly(x)|bird(x) \wedge weight(x) > \mathbf{y}]_x\Big),$$

where $Reasonable\_Weight(y)$ is a field predicate which indicates that '$\mathbf{y}$' is a number which is a reasonable weight for a bird, e.g., not negative.

Example 3.3 *Mixing universal quantification and probabilities.*

"The probability of finding a given type of animal at a zoo is given by a function, **f**, of the expense of acquiring and maintaining that type of animal:"

$$\forall x\Big(animal\_type(x) \rightarrow [at(x,y)|zoo(y)]_y = \mathbf{f}(expense(x))\Big),$$

where *expense* is a measuring function symbol and **f** is a field function symbol. Here *expense* may be a function which can be calculated through other information in the knowledge base, e.g.,

$$\forall x(expense(x) = weight(x) \times 100 + initial\_cost(x)).$$

Also, **f** could be declared to be non-decreasing:

$$\forall \mathbf{xy}(\mathbf{x} > \mathbf{y} \rightarrow \mathbf{f}(\mathbf{x}) > \mathbf{f}(\mathbf{y})).$$

Example 3.4 *Knowledge of independence (in most other systems this would be meta-knowledge):*

The canonical tri-functional expression of independence (see Pearl [56]) "The properties P and Q are independent given R:"

$$[P(x) \wedge Q(x)|R(x)]_x = [P(x)|R(x)]_x \times [Q(x)|R(x)]_x.$$

Example 3.5 *Notions from Statistics.*

1. "A sequence of ten tosses of a fair coin will land heads with a frequency between 45–55% with greater than 95% probability:"

   $$[frequency\_heads(x) \in (0.45, 0.55) | sequence\_of\_tosses(x)]_x > 0.95.$$

   Here the domain contains a set of objects, $sequence\_of\_tosses(x)$, with each member representing a sequence of ten coin tosses of a fair coin, and a measuring function, *frequency_heads*, which maps each sequence of tosses to a number in the closed interval [0,1], a number which represents the relative frequency of heads in that sequence.

2. Other notions from Statistics, e.g, "The height of adult males (humans) is normally distributed with mean 177cm and standard deviation 13cm:"

   $$\forall \mathbf{x}\mathbf{y}\Big([height(x) \in (\mathbf{x}, \mathbf{y}) | Adult\_male(x)]_x = \mathbf{normal}(\mathbf{x}, \mathbf{y}, 177, 13)\Big)$$

   Here **normal** is a field function which, given an interval $(\mathbf{x}, \mathbf{y})$, a mean, and a standard deviation, returns an approximation of the integral (approximation since we don't necessarily have access to real numbers) of a normal distribution, with specified mean and standard deviation, over the given interval.

Chapter 4

Deductive Proof Theory

This chapter provides a deductive proof theory for **Lp**. The proof theory consists of a set of axioms and rules of inference, and is shown to be both sound and complete. In fact, the proof theory for **Lp** is essentially the same as the proof theory for normal first order logic[1], the major change being in the set of axioms. Two new sets of axioms must be introduced, one set to deal with the logic of the probability function, and another set to define the logic of the field $\mathcal{F}$. There are, however, some technical difficulties arising from the probability function.

One difficulty arises from the fact that when probability terms are formed by rule **T2** (section 3.2) all of the variables $x_i \in \vec{x}$, which appear in the formula α, are bound by the probability term former. That is, their semantic interpretation is altered, as specified by the rule of interpretation **T2** (section 3.5). This creates a difficulty with those formulas which also contain other quantifiers, a difficulty which is similar to the difficulty arising from nested quantifiers in ordinary first order logic.

---

[1] A reference for all discussions of first order logic in this chapter is the textbook *A Course in Mathematical Logic* by Bell and Machover [3]

One of the rules of inference in first order logic allows terms to be substituted for the variable bound by the universal quantifier. For example, in first order logic it is valid to infer the sentence "$Man(Socrates) \rightarrow Mortal(Socrates)$" given the sentence "$\forall x(Man(x) \rightarrow Mortal(x))$". Here the term *Socrates* has been substituted for the bound variable x. When first order quantifiers are nested care must be used to avoid invalid conclusions. For example, in the formula $\forall x P(x) \rightarrow \exists x Q(x)$ a term t substituted for the first (universally) quantified x cannot be substituted for the second x; the second x is in the scope of a distinct quantifier. Such a substitution would lead to the erroneous conclusion $P(t) \rightarrow Q(t)$. In general, terms can only be substituted for the free occurrences of a variable in a formula.

Another difficulty arises from the fact that the term t may itself contain variables (especially in **Lp**, where the probability terms can contain arbitrary open formulas). When such a term is substituted into a formula its variables may be accidently captured by other quantifiers in the formula. For example, in the formula $\forall x \exists y P(x) \wedge Q(y)$ if the term $f(y)$ is substituted for the variable x the formula $\exists y P(f(y)) \wedge Q(y)$ results, where the y in $f(y)$ has been captured by the existential quantifier. This formula cannot be validly inferred from the previous formula.

Since the probability terms bind variables, these two difficulties arise in the interaction of the probability terms with the ordinary quantifiers $\forall$ and $\exists$. These difficulties are dealt with, as in first order logic, by precise definitions which specify when a given variable is free in a given formula. Substitution of terms for variables is then defined in such a way that only free variables are affected. The problem of accidental capture is overcome by developing rules for renaming quantified variables. These rules transform formulas to new formulas which are identical in their semantic meaning and in which there is no possibility of accidently capturing any of the variables in the term to be substituted in.

The fact that the probability function generates terms from formulas creates another difficulty. Most of the theorems of first order logic are proved by induction on the formulas of the logic. With **Lp** these theorems must be proved by simultaneous induction on both the formulas and the terms of the logic.

The development of the proof theory consists of two parts. First we define substitution to suit the requirements of **Lp**. After this, we present the axioms and rules of inference which make up the deductive proof theory of **Lp**. This proof theory is shown to be sound and complete.

4.1 Substitution

In this section α and β are used to refer **either** to terms or formulas of **Lp**. We begin with a definition of when a given instance of a variable in a formula is free and when it is bound. This definition is central to the development of a sound notion of substitution.

Definition 4.1.1 *A given occurrence of a variable x in a formula or term α is* free *iff it is not* bound *in α. Furthermore:*

1. If α is the variable x then x is *free* in α.
2. If $\alpha = ft_1 \ldots t_n$ or $\alpha = Pt_1 \ldots t_n$, then a given occurrence of x in α is *free* in α iff it is free in the term t_i in which it occurs.
3. If $\alpha = \neg\beta$ then a given occurrence of x in α is *free* in α iff that occurrence is free in β.
4. If $\alpha = \beta \wedge \delta$ then a given occurrence of x in α is *free* in α iff that occurrence is a free occurrence of x in β or δ.
5. If $\alpha = \forall x\beta$ then every occurrence of x in α is *bound* in α, but if $\alpha = \forall y\beta$, where y is a variable other than x, then a given occurrence of x in α is *free* in α iff that occurrence is free in β.
6. If $\alpha = [\beta]_{\vec{x}}$, and $x = x_i \in \vec{x}$ (for some i), then every occurrence of x in α is *bound* in α. Otherwise, a given occurrence of x in α is *free* in

α iff that occurrence is free in β.

We say x is *free in* α if x has at least one free occurrence in α. The *free variables* of α are all those variables which are free in α. The next theorem shows that it is only the free variables of a formula or term which can alter its meaning, once we have fixed on a specific **Lp**-Structure.

Theorem 4.1.2 *Let σ and τ be interpretations with the same underlying structure $\mathcal{M}$ which agree on every free variable of α. Then*

$$\alpha^\sigma = \alpha^\tau.$$

Proof The theorem is proved by induction on the length of α. The claim is obvious if α is a single variable or constant. If $\alpha = ft_1 \ldots t_n$ then since σ and τ have the same underlying structure $\alpha^\sigma = f^\sigma t_1^\sigma \ldots t_n^\sigma = f^\tau t_1^\sigma \ldots t_n^\sigma$. The free variables in t_i are a subset of the free variables in α; therefore, by the inductive hypothesis, $f^\tau t_1^\sigma \ldots t_n^\sigma = f^\tau t_1^\tau \ldots t_n^\tau$. The last term is α^τ. A similar argument holds when $\alpha = Pt_1 \ldots t_n$.

If $\alpha = \neg\beta$ or $\alpha = \beta \wedge \delta$ the free variables in β and δ will be a subset of the free variables in α. So by induction, $\beta^\sigma = \beta^\tau$ also $\delta^\sigma = \delta^\tau$. The claim now follows easily from the semantic definition.

If $\alpha = \forall x\beta$ then the free variables of β are exactly the free variables of α as well as possibly x. By the semantic definition

$$\alpha^\sigma = \top \quad \text{iff} \quad \beta^{\sigma(x/a)} = \top \quad \text{for all} \quad a \in \mathcal{O}.$$

Since $\sigma(x/a)$ and $\tau(x/a)$ agree on all the free variables of β, we have from the inductive hypothesis

$$\beta^{\sigma(x/a)} = \beta^{\tau(x/a)}.$$

Thus, by the semantic definition,

$$\alpha^{\sigma} = \top \quad \text{iff} \quad \alpha^{\tau} = \top.$$

If $\alpha = [\beta]_{\vec{x}}$ then the free variables of β are exactly the free variables of α as well as possibly the variables $x_i \in \vec{x}$. By the semantic definition,

$$[\beta]^{\sigma}_{\vec{x}} = \mu_n\{\vec{a} | \beta^{\sigma(\vec{x}/\vec{a})} = \top\}.$$

Since $\sigma(\vec{x}/\vec{a})$ agrees with $\tau(\vec{x}/\vec{a})$ on all of the free variables of β, the inductive hypothesis gives

$$\beta^{\sigma(\vec{x}/\vec{a})} = \beta^{\tau(\vec{x}/\vec{a})}.$$

Thus

$$\vec{c} \in \{\vec{a} | \beta^{\sigma(\vec{x}/\vec{a})} = \top\} \quad \text{iff} \quad \vec{c} \in \{\vec{a} | \beta^{\tau(\vec{x}/\vec{a})} = \top\},$$

and

$$[\beta]^{\sigma}_{\vec{x}} = \mu_n\{\vec{a} | \beta^{\tau(\vec{x}/\vec{a})} = \top\} = [\beta]^{\tau}_{\vec{x}}.$$

■

A formula α which has no free variables is called a *sentence* or a *closed* formula. This theorem implies that the truth value of a sentence, α^{σ}, depends only on the underlying structure $\mathcal{M}$. This allows a definition of structure (model) satisfaction.

Definition 4.1.3 *An* **Lp**-*structure* $\mathcal{M}$ satisfies *a sentence* α, *written* $\mathcal{M} \models \alpha$, *if* $\alpha^{\sigma} = \top$ *for all interpretations,* σ, *whose underlying structure is* $\mathcal{M}$. *More generally, an interpretation* σ satisfies *a formula* α *(set of formulas* Φ*) if* $\alpha^{\sigma} = \top$ *(*$\beta^{\sigma} = \top$ *for every* $\beta \in \Phi$*), written* $\sigma \models \alpha$ *(*$\sigma \models \Phi$*). Finally, a set of formulas* Φ entails *a formula* α *(written* $\Phi \models \alpha$*) if every interpretation which satisfies* Φ *also satisfies* α.

Now we can give a preliminary definition of substitution. This definition works when there is no possibility of an accidental capture of the variables present in the term to be substituted in. When such a condition holds we say that the term is *free for* (free to be substituted for) the variable in question. The following definition gives the precise conditions under which a term, t, is free for a given variable, x, in a formula, α. It also specifies the exact form of the new formula, denoted by $\alpha(x/t)$, which is the result of performing the substitution. Of course, for substitution to make sense semantically the term t and the variable x must be of the same type; it is assumed that they are, throughout this section.

Definition 4.1.4 *Given a term, t, a variable of the same type, x, and a formula or term, α, t is* free for x in α *under the following conditions:*

1. If α is a variable or constant then t is *free for x in α*. If $\alpha = x$ then $\alpha(x/t) = t$, otherwise, $\alpha(x/t) = \alpha$.

2. If $\alpha = ft_1 \dots t_n$ or $\alpha = Pt_1 \dots t_n$, then t is *free for x in α* iff t is free for x in every term t_i. And $\alpha(x/t)$ is defined as $ft_1(x/t) \dots t_n(x/t)$ or $Pt_1(x/t) \dots t_n(x/t)$.

3. If $\alpha = \neg\beta$ then t is *free for x in α* iff t is free for x in β. In this case, $\alpha(x/t)$ is defined as $\neg(\beta(x/t))$.

4. If $\alpha = \beta \wedge \delta$ then t is *free for x in α* iff t is free for x in both β and δ; $\alpha(x/t)$ is defined as $\beta(x/t) \wedge \delta(x/t)$.

5. If $\alpha = \forall y\beta$ then t is *free for x in α* iff one of the following conditions hold:

   (i) x is not free in α (as is the case when $x = y$); $\alpha(x/t)$ is defined as α.

   (ii) x is free in α, t is free for x in β, and y is not free in t; $\alpha(x/t)$ is defined as $\forall y(\beta(x/t))$.

6. If $\alpha = [\beta]_{\vec{y}}$ then t is *free for x in α* iff one of the following conditions holds:

   (i) x is not free in α (as in the case when $x \in \vec{y}$); $\alpha(x/t)$ is defined as α.

   (ii) x is free in α, t is free for x in β, and no $y_i \in \vec{y}$ is free in t; $\alpha(x/t)$ is defined as $[\beta(x/t)]_{\vec{y}}$.

The next theorem clarifies the semantic behavior of legal substitutions as defined above.

Theorem 4.1.5 *If t is free for x in α, then for every interpretation σ*

$$\alpha(x/t)^{\sigma} = \alpha^{\sigma(x/t')} \qquad \textit{where } t' = t^{\sigma}.$$

Proof By induction on the length of α. If α is a variable or constant not equal to x then $\alpha = \alpha(x/t)$. If $\alpha = x$ then $\alpha(x/t)^{\sigma} = t^{\sigma} = t' = x^{\sigma(x/t')} = \alpha^{\sigma(x/t')}$.

If $\alpha = ft_1 \ldots t_n$ (or $Pt_1 \ldots t_n$) then $\alpha(x/t)^{\sigma} = f^{\sigma} t_1(x/t)^{\sigma} \ldots t_n(x/t)^{\sigma} = f^{\sigma(x/t')} t_1^{\sigma(x/t')} \ldots t_n^{\sigma(x/t')}$ by induction.

If $\alpha = \neg\beta$ or $\alpha = \beta \wedge \delta$ the claim follows easily from the inductive hypothesis and the semantic definition.

If $\alpha = \forall y \beta$ then one of two conditions holds:

(i) x is not free in α, in which case $\alpha(x/t) = \alpha$ and the conditions of theorem 4.1.2 hold. That is, $\alpha^{\sigma} = \alpha^{\sigma(x/t')}$, so $\alpha(x/t)^{\sigma} = \alpha^{\sigma} = \alpha^{\sigma(x/t')}$.

(ii) x is free in α, t is free for x in β, and y is not free in t, in which case $\alpha(x/t) = \forall y \beta(x/t)$. By the semantic definition[2]

$$(\forall y \beta(x/t))^{\sigma} = \top \quad \text{iff} \quad \beta(x/t)^{\sigma(y/a)} = \top \quad \text{for all} \quad a \in \mathcal{O}. \tag{1}$$

---

[2] If y is a field variable then the set of objects $\mathcal{O}$ should be replaced by the field $\mathcal{F}$ in this part of the proof.

Since t is free for x in β, we have by the inductive hypothesis

$$\beta(x/t)^{\sigma(y/a)} = \beta^{\sigma(y/a)(x/t'')} \quad \text{where} \quad t'' = t^{\sigma(y/a)}.$$

y is not free in t so, by theorem 4.1.2, $t'' = t^{\sigma(y/a)} = t^{\sigma} = t'$. Also, since x and y are distinct (otherwise x would not be free in α), $\sigma(y/a)(x/t') = \sigma(x/t')(y/a)$. Hence,

$$\beta(x/t)^{\sigma(y/a)} = \beta^{\sigma(x/t')(y/a)}. \qquad (3)$$

By the semantic definition

$$\beta^{\sigma(x/t')(y/a)} = \top \quad \text{for all} \quad a \in \mathcal{O} \quad \text{iff} \quad (\forall y \beta)^{\sigma(x/t')} = \top.$$

This, along with (1) and (3), gives the desired result.

Similarly, if $\alpha = [\beta]_{\vec{y}}$ then one of two conditions holds:

(i) x is not free in α. This case is identical to case (i) above.

(ii) x is free in α, t is free for x in β, and no $y_i \in y$ is free in t. In this case $\alpha(x/t) = [\beta(x/t)]_{\vec{y}}$. By the semantic definition

$$[\beta(x/t)]_{\vec{y}}^{\sigma} = \mu^{n}\{\vec{a} | \beta(x/t)^{\sigma(\vec{y}/\vec{a})} = \top\}.$$

As t is free for x in β, the inductive hypothesis yields

$$\beta(x/t)^{\sigma(\vec{y}/\vec{a})} = \beta^{\sigma(\vec{y}/\vec{a})(x/t'')} \qquad \text{where} \quad t'' = t^{\sigma(\vec{y}/\vec{a})}.$$

Since no $y_i \in \vec{y}$ is free in t, theorem 4.1.2 implies $t'' = t^{\sigma(\vec{y}/\vec{a})} = t^{\sigma} = t'$. Also, x is not equal to any $y_i \in \vec{y}$ so $\sigma(\vec{y}/\vec{a})(x/t') = \sigma(x/t')(\vec{y}/\vec{a})$. Hence,

$$\beta(x/t)^{\sigma(\vec{y}/\vec{a})} = \beta^{\sigma(x/t')(\vec{y}/\vec{a})}$$

Therefore,

$$\vec{c} \in \{\vec{a} | \beta(x/t)^{\sigma(\vec{y}/\vec{a})} = \top\} \quad \text{iff} \quad \vec{c} \in \{\vec{a} | \beta^{\sigma(x/t')(\vec{y}/\vec{a})} = \top\}.$$

So, from the semantic definition, $[\beta(x/t)]_{\vec{y}}^{\sigma} = [\beta]_{\vec{y}}^{\sigma(x/t')}$ as claimed.

■

To deal with substitution when an accidental capture would occur it is necessary to rename quantified variables in formulas. The next definition gives rules for renaming which preserve meaning. They expand the standard first order definitions by allowing the renaming of variables bound by the probability function.

Definition 4.1.6 *The* variants *of α are the following:*

(a) α is its own *variant.*

(b) Any *direct variant* of α, as defined below, is a variant of α:

1. If α is a variable or constant then it has no direct variants.

2. If $\alpha = ft_1 \ldots t_n$ or $\alpha = Pt_1 \ldots t_n$, then the ***direct variants*** of α are all terms (formulas) of the form $ft'_1 \ldots t'_n$ $(Pt'_1 \ldots t'_n)$, where $t'_i = t_i$ or t'_i is a direct variant of t_i and at least one of the $t'_i \neq t_i$.

3. If $\alpha = \neg\beta$ then the ***direct variants*** of α are all formulas of the form $\neg(\beta')$, where β' is a direct variant of β.

4. If $\alpha = \beta \wedge \delta$ then the ***direct variants*** of α are all formulas of the form $\beta' \wedge \delta'$, where β' and δ' are direct variants of β and δ respectively.

5. If $\alpha = \forall x\beta$ then the ***direct variants*** of α are all formulas of the form $\forall x\beta'$, where β' is a direct variant of β. As well as, all formulas $\forall z(\beta'(x/z))$ formed from $\forall x\beta'$, where z is a variable of the same type as x but not equal to x, z is not free in β', and z is free for x in β' (e.g., these conditions are met if z does not occur in β').

6. If $\alpha = [\beta]_{\vec{x}}$ then the ***direct variants*** of α are all terms of the form $[\beta']_{\vec{x}}$, where β' is a ***direct variant*** of β. As well as, all terms $[\beta'(x_i/z)]_{\vec{x}(x_i/z)}$ formed from $[\beta']_{\vec{x}}$, where z is an object variable not equal to any $x_i \in \vec{x}$, which is not free in β' but is free for x_i in β' (e.g., if z does not occur in β') and $\vec{x}(x_i/z)$ is a vector of object variables identical to $\vec{x}$ except that z has been substituted for the *i-th* variable x_i.

(c) Any formula (term) α' which results from a sequence of direct variations of α is a *variant* of α. That is, if $\alpha' = \alpha_n$ and $\alpha = \alpha_1$, where, in the sequence $(\alpha_1, \ldots, \alpha_n)$, α_{i+1} is a direct variant of α_i, then α' is a variant of α.

We can prove that variable renaming in this manner preserves meaning.

Theorem 4.1.7 *If α' is a variant of α then for every interpretation function σ*

$$\alpha'^{\sigma} = \alpha^{\sigma}.$$

Furthermore, the underlying sets defined by two variant probability terms are identical. That is, if $\alpha' = [\beta']_{\vec{y}}$ and $\alpha = [\beta]_{\vec{x}}$, then

$$\{\vec{a} | \beta'^{\sigma(\vec{y}/\vec{a})} = \top\} = \{\vec{a} | \beta^{\sigma(\vec{x}/\vec{a})} = \top\}.$$

Proof The theorem is proved by induction on the number of direct variations applied to α to yield α'.

0) If α' is equal to α (i.e. zero variations applied) the claim is obvious.

1) If α' is the result of one direct variation of α then we prove the claim by induction of the length of α. First, α cannot be a variable or constant as then it would not have any direct variants. If $\alpha = ft_1 \ldots t_n$ or $\alpha = Pt_1 \ldots t_n$, then $\alpha' = ft'_1 \ldots t'_n$ $(Pt'_1 \ldots t'_n)$, where the t'_i are either direct variants of the corresponding t_i or are equal to t_i. In this case, the claim follows

directly from the inductive hypothesis and the semantic definition, similarly if $\alpha = \neg\beta$ or $\alpha = \beta \wedge \delta$.

If $\alpha = \forall x\beta$ then two cases arise:

(i) $\alpha' = \forall x\beta'$, where β' is a direct variant of β. By induction $\beta'^{\sigma(x/u)} = \beta^{\sigma(x/u)}$, so by the semantic definition, $(\forall x\beta')^{\sigma} = (\forall x\beta)^{\sigma}$.

(ii) $\alpha' = \forall z(\beta'(x/z))$, where β' is a direct variant of β, z is not free in β', and z is free for x in β'. Since z is free for x in β' we have $\beta'(x/z)^{\tau} = \beta'^{\tau(x/z')}$ $(z' = z^{\tau})$ for any interpretation τ, by theorem 4.1.5. Therefore,

$$\beta'(x/z)^{\sigma(z/u)} = \beta'^{\sigma(z/u)(x/z')} \quad \text{where} \quad z' = z^{\sigma(z/u)}. \tag{1}$$

By the definition of $\sigma(z/u)$, $z' = u$. Furthermore, since z is not free in β', $\sigma(z/u)(x/z')$ and $\sigma(x/z')$ agree on all of the free variables of β'; thus by theorem 4.1.2, $\beta'^{\sigma(z/u)(x/z')} = \beta'^{\sigma(x/z')}$. Hence,

$$\beta'^{\sigma(z/u)(x/z')} = \beta'^{\sigma(x/z')} = \beta'^{\sigma(x/u)}. \tag{2}$$

By the induction hypothesis

$$\beta'^{\sigma(x/u)} = \beta^{\sigma(x/u)}.$$

The result follows from this along with (2), (1), and the semantic definition.

If $\alpha = [\beta]_x$ then two cases also arise.

(i) $\alpha' = [\beta']_{\vec{x}}$, where β' is a direct variant of β. By induction $\beta'^{\sigma} = \beta^{\sigma}$ for any interpretation σ; therefore, $\vec{c} \in \{\vec{a} | \beta'^{\sigma(\vec{x}/\vec{a})} = \top\}$ iff $\vec{c} \in \{\vec{a} | \beta^{\sigma(\vec{x}/\vec{a})} = \top\}$. So the claim follows from the semantic definition.

(ii) $\alpha' = [\beta'(x_i/z)]_{\vec{x}(x_i/z)}$, where β' is a direct variant of β, z is a variable not free in β', but free for x_i in β'. Since z is free for x_i we have from theorem 4.1.5

$$\beta'(x_i/z)^{\sigma(\vec{x}(x_i/z)/\vec{a})} = \beta'^{\sigma(\vec{x}(x_i/z)/\vec{a})(x_i/z')}, \qquad z' = z^{\sigma(\vec{x}(x_i/z)/\vec{a})}.$$

Since z is not free in β', $\sigma(\vec{x}(x_i/z)/\vec{a})$ and

$$\sigma(\langle x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\rangle / \langle a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n\rangle)$$

agree on all the free variables of β'. So theorem 4.1.2 implies

$$\beta'^{\sigma(\vec{x}(x_i/z)/\vec{a})(x_i/z')} = \beta'^{\sigma(\langle x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n\rangle/\langle a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_n\rangle)(x_i/z')}.$$

Also, by definition $z' = a_i$, hence,

$$\beta'^{\sigma(\langle x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n\rangle/\langle a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_n\rangle)(x_i/z')} = \beta'^{\sigma(\vec{x}/\vec{a})}.$$

By the induction hypothesis $\beta'^{\sigma(\vec{x}/\vec{a})} = \beta^{\sigma(\vec{x}/\vec{a})}$; thus,

$$\vec{c} \in \left\{\vec{a} | \beta'(x_i/z)^{\sigma(\vec{x}(x_i/z)/\vec{a})} = \top\right\} \quad \text{iff} \quad \vec{c} \in \left\{\vec{a} | \beta^{\sigma(\vec{x}/\vec{a})} = \top\right\}.$$

This proves the stronger condition satisfied by variant probability terms. The first claim of the theorem follows from this equivalence and the semantic definition.

n) If α' is the result of n direct variations of α, then α' is a direct variant of α_{n-1}, where α_{n-1} is the result of $n-1$ direct variations of α. By induction, $\alpha^{\sigma}_{n-1} = \alpha^{\sigma}$ also $\alpha^{\sigma}_{n-1} = \alpha'^{\sigma}$, so, it is obvious that $\alpha'^{\sigma} = \alpha^{\sigma}$. ■

This definition and theorem allow a final definition of substitution. In the final definition there is no condition of 'free for'. The definition works by first specifying a variant formula, a variant in which the given term is free for the given variable. Once this variant is formed substitution can occur, just as in definition 4.1.4. The definition is an extension of definition 4.1.4; when the given term is already free for the given variable the variant is the original formula itself. As before, $\alpha(x/t)$ denotes the new formula which results from the substitution of the term t for the variable x in the formula α.

Definition 4.1.8 (Substitution) *For given x, t, and α define $\alpha(x/t)$ to be $\alpha'(x/t)$, where $\alpha'(x/t)$ is defined according to definition 4.1.4 and α' is defined as follows:*

1. If α is a constant or variable then $\alpha' = \alpha$.

2. If $\alpha = ft_1 \ldots t_n$ or $Pt_1 \ldots t_n$ then $\alpha' = ft'_1 \ldots t'_n$ $(Pt'_1 \ldots t'_n)$.

3. If $\alpha = \neg\beta$ then $\alpha' = \neg(\beta')$.

4. If $\alpha = \beta \wedge \delta$ then $\alpha' = \beta' \wedge \delta'$.

5. If $\alpha = \forall y \beta$ there are three cases:

   a) If x is not free in α then $\alpha' = \alpha$.

   b) If x is free in α and y is not free in t, then $\alpha' = \forall y \beta'$.

   c) If x is free in α and y is free in t, then $\alpha' = \forall z(\beta'(y/z))$. Where z is the first variable (in some fixed but arbitrary enumeration of the variables) of the same type as y which does not occur in either t or β'.

6. If $\alpha = [\beta]_{\vec{y}}$ again there are three cases:

   a) If x is not free in α then $\alpha' = \alpha$.

   b) If x is free in α and no $y_i \in \vec{y}$ is free in t, then $\alpha' = [\beta']_{\vec{y}}$.

   c) If x is free in α and some subset of $y_i \in \vec{y}$ (call it $\vec{y}' = \{y'_1, \ldots, y'_m\}$) are free in t, then $\alpha' = [\beta'(\vec{y}'/\vec{z})]_{\vec{y}(\vec{y}'/\vec{z})}$. Where $\vec{z} = \{z_1, \ldots, z_m\}$ is a set of new object variables which do not appear in either t or β'. The substitution of z_i for y'_i is done one at a time, forming a sequence of m direct variants.

The next theorem shows that the general form of substitution has the same semantic behaviour as the preliminary form.

Theorem 4.1.9 *For all α, t, x and interpretations σ*

$$\alpha(x/t)^{\sigma} = \alpha^{\sigma(x/t')} \qquad \textit{where } t' = t^{\sigma}.$$

Proof By definition 4.1.8, $\alpha(x/t) = \alpha'(x/t)$, where α' is a variant of α and t is free for x in α'. By theorem 4.1.5, $\alpha'(x/t)^{\sigma} = \alpha'^{\sigma(x/t')}$, and by theorem 4.1.7, $\alpha'^{\sigma(x/t')} = \alpha^{\sigma(x/t')}$. ■

The results of this section allow us to prove that the subsets of $\mathcal{O}^n$ defined by the formulas of **Lp** forms a field of subsets. This fact will be used later in the proof of completeness.

Theorem 4.1.10 *The set of subsets of* $\mathcal{O}^n$ *defined by the formulas of* **Lp**, *is a field of subsets.*

Proof Let A and B be two subsets of $\mathcal{O}^n$ defined by formulas of **Lp**, i.e., $A = \{\vec{a} | \alpha^{\sigma(\vec{x}/\vec{a})} = \top\}$ and $B = \{\vec{b} | \beta^{\sigma(\vec{y}/\vec{b})} = \top\}$. By definition 4.1.6, there exists two variants of $[\alpha]_{\vec{x}}$ and $[\beta]_{\vec{y}}$, $[\alpha']_{\vec{z}}$ and $[\beta']_{\vec{z}}$, formed by substituting all the variables $x_i \in \vec{x}$ in α and all the variables $y_i \in \vec{y}$ in β by a new set of variables $\langle z_1, \ldots, z_n \rangle$ which do not appear in α or β. By theorem 4.1.7, $A' = \{\vec{c} \,| \alpha'^{\sigma(\vec{z}/\vec{z})} = \top\} = A$, and $B' = \{\vec{c} \,| \beta'^{\sigma(\vec{z}/\vec{c})} = \top\} = B$; thus, $A \cap B = A' \cap B' = \{\vec{c} \,| (\beta' \wedge \alpha')^{\sigma(\vec{z}/\vec{c})} = \top\}$. That is, the intersection of A and B is definable by a formula of **Lp**. Similarly, for A, as defined above, by the semantic definition, $\alpha^{\sigma(\vec{x}/\vec{a})} = \top$ iff $\neg\alpha^{\sigma(\vec{x}/\vec{a})} = \bot$. Thus, $\vec{a} \in A$ iff $\vec{a} \notin A' = \{\vec{a} | \neg\alpha^{\sigma(\vec{x}/\vec{a})} = \top\}$. That is, A' is the complement of A with respect to $\mathcal{O}^n$, and is definable by a formula of **Lp**. Hence, the set of subsets of $\mathcal{O}^n$ definable by formulas of **Lp** is closed under intersections and complementations. Finally, if we take the term $[\alpha \wedge \neg\alpha]_{\vec{x}}$ the set $A =$

$\{\vec{a} | (\alpha \wedge \neg \alpha)^{\sigma(\vec{x}/\vec{a})} = \top\}$ is empty; thus the empty set is definable by a formula of **Lp**. Q.e.d. ■

4.2 Proof Theory

This section gives a proof theory for **Lp**. The proof theory consists of a set of axioms and rules of inference, and it is shown to be both sound and complete. There are, in addition to the normal first order axioms, two new sets of axioms. One set of axioms defines the logic of the probability terms, and the other set defines the logic of the field $\mathcal{F}$.

In this section α, β, etc., will usually be used to represent formulas, not formulas or terms, as was the common usage in the previous section. It will be explicitly stated when they may also refer to terms.

4.2.1 Axioms and Rules of Inference

First the axioms and rules of inference (actually there is only one) for the proof theory are presented.

If α is a formula of **Lp** then a *generalization* of α is any formula of the form $\forall x_1 \ldots \forall x_n \alpha$, where $\{x_1, \ldots, x_n\}$ is a set of not necessarily distinct variables of either type.

First order Axioms All the axioms of the Predicate Calculus.

PC1a) $\alpha \rightarrow \beta \rightarrow \alpha$.

PC1b) $(\alpha \to \beta \to \delta) \to (\alpha \to \beta) \to \alpha \to \delta$.

PC1c) $(\neg\alpha \to \beta) \to (\neg\alpha \to \neg\beta) \to \alpha$.

PC2) $\forall x(\alpha \to \beta) \to \forall x\alpha \to \forall x\beta$.

PC3) $\alpha \to \forall x\alpha$,

where x is not free in α.

PC4) $\forall x\alpha \to \alpha(x/t)$,

where t is any term, of the same type as x, free for x in α, and $\alpha(x/t)$ is defined according to definition 4.1.4.

EQ5) $t = t$,

where t is any term.

EQ6) $t_1{=}t_{n+1} \to \cdots \to t_n{=}t_{2n} \to ft_1 \ldots t_n{=}ft_{n+1} \ldots t_{2n}$,

where f is any n-ary function symbol and $t_1, \ldots, t_{2n}$ are terms of a compatible type.

EQ7) $t_1{=}t_{n+1} \to \cdots \to t_n{=}t_{2n} \to Pt_1 \ldots t_n {\to} Pt_{n+1} \ldots t_{2n}$,

where P is any n-ary predicate symbol and $t_1, \ldots, t_{2n}$ are terms of the same type.

Field Axioms All of the axioms of a totally ordered field (see, for example, MacLane [45]). Here all variables are field variables and they are all *universally quantified*, unless the existential quantifier is used.

F1) $x + (y + z) = (x + y) + z$

F2) $x + 0 = x$

F3) $\exists y(x + y = 0)$

F4) $x + y = y + x$

F5) $x \times 1 = x$

F6) $x \times (y \times z) = (x \times y) \times z$

F7) $x \times y = y \times x$

F8) $x \times (y + z) = (x \times y) + (x \times z)$

F9) $1 \geq 0 \wedge \neg(1 = 0)$

F10) $x \neq 0 \rightarrow \exists y(y \times x = 1)$

F11) $(x \geq y \wedge y \geq z) \rightarrow x \geq z$

F12) $(x \geq y \wedge y \geq x) \rightarrow x = y$

F13) $x \geq x$

F14) $x \geq y \vee y \geq x$

F15) $x \geq y \rightarrow x + z \geq y + z$

F16) $(x \geq y \wedge z \geq 0) \rightarrow x \times z \geq y \times z$

Probability Function Axioms

P1) $\forall x_1 \ldots \forall x_n \alpha \rightarrow [\alpha]_{\vec{x}} = 1$,

where $\vec{x} = \langle x_1, \ldots, x_n \rangle$ and every x_i is an object variable.

P2) $[\alpha]_{\vec{x}} \geq 0$.

P3) $[\alpha]_{\vec{x}} + [\neg\alpha]_{\vec{x}} = 1$.

P4) $[\alpha]_{\vec{x}} + [\beta]_{\vec{x}} \geq [\alpha \vee \beta]_{\vec{x}}$.

P5) $[\alpha \wedge \beta]_{\vec{x}} = 0 \rightarrow [\alpha]_{\vec{x}} + [\beta]_{\vec{x}} = [\alpha \vee \beta]_{\vec{x}}$.

P6) $[\alpha]_{\vec{x}} = [\alpha(x_i/z)]_{\vec{x}(x_i/z)}$,

where z is an object variable which is free for x_i in α, z is not free in α, and $\vec{x}(x_i/z)$ is a new vector of object variables:

$\langle x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n \rangle$.

P7) $\forall \mathbf{z}_1\mathbf{z}_2[[\alpha]_{\vec{x}} = \mathbf{z}_1]_{\vec{y}} = \mathbf{z}_2 \rightarrow ([\alpha]_{\langle \vec{x}, \vec{y} \rangle} \geq \mathbf{z}_1 \times \mathbf{z}_2)$.

P8) (Required for infinite domains only)

$[\alpha]_{\vec{x}} = [\alpha]_{\pi(\vec{x})}$,

where π is any permutation of $\{1, \ldots, n\}$, and $\pi(\vec{x})$ is the permuted vector $\vec{x}$, i.e., $\pi(\vec{x}) = \langle x_{\pi(1)}, \ldots, x_{\pi(n)} \rangle$.

P9) (Required for infinite domains only)

$$\forall \mathbf{z}_1\mathbf{z}_2\Big(\Big[(\mathbf{z}_1 \leq [\alpha|\beta]_{\vec{x}} \leq \mathbf{z}_2) \wedge \beta\Big]_{\langle \vec{x}, \vec{y} \rangle} \neq 0$$

$$\rightarrow \mathbf{z}_1 \leq \Big[\alpha \big| (\mathbf{z}_1 \leq [\alpha|\beta]_{\bar{x}} \leq \mathbf{z}_2) \wedge \beta\Big]_{\langle \bar{x}, \bar{y}\rangle} \leq \mathbf{z}_2\Big).$$

Generalization

G1) All generalizations of the preceding axioms.

Rule of inference The only rule of inference is *modus ponens*, i.e.,

R1) From $\{\alpha, \alpha \rightarrow \beta\}$ infer β.

4.2.2 Deductions

This section defines the notion of a special sequence of formulas called a deduction, and notes some of its properties.

Definition 4.2.1 *Let* Φ *be a set of* Lp *formulas. A* deduction from Φ in Lp *is a finite non-empty sequence of formulas* $\phi_1, \ldots, \phi_n$ *such that for each* $k (1 \leq k \leq n)$ ϕ_k *is an axiom of* Lp, *or* $\phi_k \in \Phi$, *or* ϕ_k *is obtained by modus ponens from earlier formulas in the same sequence (i.e., there exists* $i, j \leq k$ *such that* $\phi_j = \phi_i \rightarrow \phi_k$*). The set* Φ *is called the set of hypotheses. If* Φ *is empty the deduction is called a* proof, *i.e., a proof is a deduction which just uses the axioms of* Lp. *A deduction whose last formula is* α *is called a deduction of* α. *The symbol '*$\vdash$*' is used to indicate deducibility, i.e., '*$\Phi \vdash \alpha$*' means there is a deduction of* α *from* Φ*, and '*$\vdash \alpha$*' means that there is a proof of* α.

Theorem 4.2.2 (Deduction Theorem) *Given a deduction of* β *from* $\{\Phi, \alpha\}$, *a deduction of* $\alpha \rightarrow \beta$ *can be constructed from* Φ.

Proof Standard first order proof holds. ■

Definition 4.2.3 *A variable* x *is free in a set* Φ *of formulas if* x *is free in some formula in* Φ. *Similarly,* x *is free in a deduction* D *if* x *is free in some formula in* D.

Theorem 4.2.4 *Let* x *be a variable which is not free in* Φ. *Given a deduction* D *of* α *from* Φ, *a deduction* D' *of* $\forall x\alpha$ *from* Φ *can be constructed which has the following properties:*

(i) x *is not free in* D',

(ii) *every variable free in* D' *is free in* D *as well.*

Proof Standard first order proof holds. ■

Theorem 4.2.4 has two special cases which are particularly important.

a) If $\Phi \vdash \alpha$ and x is not free in Φ, then $\Phi \vdash \forall x\alpha$.

b) If α is provable (i.e., deducible from an empty set of hypotheses) then $\forall x\alpha$ is provable for any variable x; symbolically, if $\vdash \alpha$ then $\vdash \forall x\alpha$.

4.2.3 Provable Equivalence

This section presents the important notion of provable equivalence; shows that the terms in the language are divided into equivalence classes by the '=' relation; and shows that variants (terms and formulas) possess the property of provable equivalence. These two results are important for the proof of the completeness theorem. Also, the results of this section show that the restriction that t be free for x in axiom PC4 is in fact not a restriction. The axiom was written in the restrictive form to make the proof of soundness simpler.

Definition 4.2.5 *Two formulas α and β are said to be* provably equivalent *if $\alpha \vdash \beta$ and $\beta \vdash \alpha$. Two* terms *α and β are said to be* provably equivalent *if $\vdash \alpha = \beta$.*

Theorem 4.2.6 *The following are provable in* **Lp**.

a) = is an equivalence relation. That is for any terms t_1, t_2, and t_3 of **Lp** we have:

- **(i)** $\vdash t_1 = t_1$,
- **(ii)** $\vdash t_1 = t_2 \rightarrow t_2 = t_1$,
- **(iii)** $\vdash t_1 = t_2 \wedge t_2 = t_3 \rightarrow t_1 = t_3$.

b) $([\alpha \rightarrow \beta]_{\bar{x}} = 1 \wedge [\beta \rightarrow \alpha]_{\bar{x}} = 1) \rightarrow [\alpha]_{\bar{x}} = [\beta]_{\bar{x}}$.

Proof The first proposition is needed for the completeness proof, and the second is a handy fact about the probability terms which is often used. The proofs demonstrate the nature of symbolic reasoning with the various axioms.

a) $t_1{=}t_1$ is an instance of axiom EQ5. With the predicate symbol P taken to be the equality predicate symbol '=' we have $t_1{=}t_2 \rightarrow t_1{=}t_1 \rightarrow t_2{=}t_1$ is an instance of axiom EQ7. So we have $t_1{=}t_2 \vdash t_2{=}t_1$. And, by the deduction theorem, $\vdash t_1{=}t_2 \rightarrow t_2{=}t_1$. Also, $t_2{=}t_1 \rightarrow t_2{=}t_3 \rightarrow t_2{=}t_2 \rightarrow t_1{=}t_3$ is another instance of EQ7. Since $t_1{=}t_2 \wedge t_2{=}t_3 \vdash t_1{=}t_2$ by tautologies and $t_1{=}t_2 \vdash t_2{=}t_1$, we have, through applications of modus ponens, $t_1{=}t_2 \wedge t_2{=}t_3 \vdash t_1{=}t_3$. Thus, $\vdash t_1{=}t_2 \wedge t_2{=}t_3 \rightarrow t_1{=}t_3$, by the deduction theorem.

b) We construct a deduction of $[\alpha]_{\bar{x}} = [\beta]_{\bar{x}}$ from $[\alpha \rightarrow \beta]_{\bar{x}}{=}1 \wedge [\beta \rightarrow \alpha]_{\bar{x}}{=}1$ (The axiom or rule of inference used in each step is specified at the right).

$$\big([\alpha \rightarrow \beta]_{\bar{x}} = 1 \wedge [\beta \rightarrow \alpha]_{\bar{x}} = 1\big) \rightarrow [\alpha \rightarrow \beta]_{\bar{x}} = 1 \qquad \text{(PC1)}$$

$$[\alpha \rightarrow \beta]_{\bar{x}} = 1 \wedge [\beta \rightarrow \alpha]_{\bar{x}} = 1 \qquad \text{(Hyp.)}$$

$$[\alpha \rightarrow \beta]_{\bar{x}} = 1 \qquad \text{(m.p.)}$$

$$[\neg\alpha \vee \beta]_{\bar{x}} = 1$$

$$[\neg\alpha]_{\bar{x}} + [\beta]_{\bar{x}} \geq [\neg\alpha \vee \beta]_{\bar{x}} \qquad \text{(P4)}$$

$$[\neg\alpha \vee \beta]_{\bar{x}} = 1 \rightarrow [\neg\alpha]_{\bar{x}} + [\beta]_{\bar{x}} \geq [\neg\alpha \vee \beta]_{\bar{x}} \rightarrow [\neg\alpha]_{\bar{x}} + [\beta]_{\bar{x}} \geq 1 \qquad \text{(EQ7)}$$

$$[\neg\alpha]_{\bar{x}} + [\beta]_{\bar{x}} \geq 1 \qquad \text{(m.p.)}$$

$$[\neg\alpha]_{\bar{x}} + [\alpha]_{\bar{x}} = 1 \qquad \text{(P3)}$$

$$[\neg\alpha]_{\bar{x}} + [\beta]_{\bar{x}} \geq [\neg\alpha]_{\bar{x}} + [\alpha]_{\bar{x}} \qquad \text{(EQ7)}$$

$$[\neg\alpha]_{\bar{x}} + (-)[\neg\alpha]_{\bar{x}} = 0 \qquad \text{(F3)}$$

$$[\neg\alpha]_{\bar{x}} + [\beta]_{\bar{x}} + (-)[\neg\alpha]_{\bar{x}} \geq [\alpha]_{\bar{x}} + [\alpha]_{\bar{x}} + (-)[\neg\alpha]_{\bar{x}} \qquad \text{(F15, m.p)}$$

$$[\beta]_{\bar{x}} + 0 \geq [\alpha]_{\bar{x}} + 0 \qquad \text{(F4, EQ7)}$$

$$[\beta]_{\bar{x}} \geq [\alpha]_{\bar{x}} \qquad \text{(F2, EQ7)}$$

Similarly from $[\neg\beta \vee \alpha]_{\bar{x}} = 1$ we derive

$$[\alpha]_{\bar{x}} \geq [\beta]_{\bar{x}}$$

thus

$$[\beta]_{\bar{x}} = [\alpha]_{\bar{x}} \qquad \text{(F12, m.p.)}$$

So $[\alpha \rightarrow \beta]_{\bar{x}}{=}1 \wedge [\beta \rightarrow \alpha]_{\bar{x}}{=}1 \vdash [\alpha]_{\bar{x}}{=}[\beta]_{\bar{x}}$; thus by the deduction theorem, $\vdash ([\alpha \rightarrow \beta]_{\bar{x}}{=}1 \wedge [\beta \rightarrow \alpha]_{\bar{x}}{=}1) \rightarrow [\alpha]_{\bar{x}}{=}[\beta]_{\bar{x}}$. ■

Theorem 4.2.7 *If α' is a variant of α then α' and α are provably equivalent, where α can be* either *a term or a formula.*

Proof This theorem is proved by induction on the number of direct variations applied to α to yield α'. If α' is equal to α (i.e., zero variations applied) then if α is a term, axiom EQ5 is a proof of $\alpha' = \alpha$. If α is a formula then obviously $\alpha \vdash \alpha'$, and $\alpha' \vdash \alpha$. If α' is the result of one direct variation of α we prove the claim by induction on the length of α. First, α cannot be a variable or constant as these do not have any direct variants. If $\alpha = ft_1 \ldots t_n$ or $Pt_1 \ldots t_n$, then $\alpha' = ft'_1 \ldots t'_n$ $(Pt'_1 \ldots t'_n)$, where $t'_i = t_i$ or t'_i is a direct variant of t_i. In this case we have by the inductive assumption $\vdash t_i = t'_i$ $(i = 1, \ldots, n)$. Therefore, by axiom EQ6 and n applications of modus ponens, we have $\vdash ft_1 \ldots t_n = ft'_1 \ldots t'_n$. If $\alpha = Pt_1 \ldots t_n$ then axiom EQ7 and n applications of modus ponens yields $\vdash Pt_1 \ldots t_n \rightarrow Pt'_1 \ldots t'_n$. Since $t_i = t'_i \rightarrow t'_i = t_i$ (theorem 4.2.6) we also have $\vdash Pt'_1 \ldots t'_n \rightarrow Pt_1 \ldots t_n$. If $\alpha = \neg\beta$ then $\alpha' = \neg\beta'$, and by induction, $\vdash \beta \rightarrow \beta'$ and $\vdash \beta' \rightarrow \beta$. But, $(\beta \rightarrow \beta') \rightarrow (\neg\beta' \rightarrow \neg\beta)$ is a tautology so $\vdash \neg\beta \rightarrow \neg\beta'$; similarly, $\vdash \neg\beta' \rightarrow \neg\beta$. If $\alpha = \beta \wedge \delta$ then $\alpha' = \beta' \wedge \delta'$ and $\vdash \beta \rightarrow \beta' \wedge \beta' \rightarrow \beta$ also $\vdash \delta \rightarrow \delta' \wedge \delta' \rightarrow \delta$ by induction. So by the use of tautologies $\vdash \alpha \rightarrow \alpha'$ and $\vdash \alpha' \rightarrow \alpha$.

If $\alpha = \forall x\beta$ two cases arise.

(i) $\alpha' = \forall x\beta'$, where β' is a direct variant of β. By induction, $\vdash \beta \rightarrow \beta'$. So by theorem 4.2.4, $\vdash \forall x(\beta' \rightarrow \beta)$. Using axiom PC2 and modus ponens we have $\vdash \forall x\beta \rightarrow \forall x\beta'$. Similarly, from $\vdash \beta' \rightarrow \beta$ we deduce $\vdash \forall x\beta' \rightarrow \forall x\beta$.

(ii) $\alpha' = \forall z(\beta'(x/z))$, where β' is a direct variant of β, z is not free in β', and z is free for x in β'. Since z is free for x in β', we have $\forall x\beta' \rightarrow \beta'(x/z)$ as an instance of axiom PC4. Thus, $\forall x\beta' \vdash \beta'(x/z)$. Also, z is not free in β' nor in $\forall x\beta'$, so by theorem 4.2.4, $\forall x\beta' \vdash \forall z\beta'(x/z)$. By the previous result $\forall x\beta \vdash \forall x\beta'$, so, $\forall x\beta \vdash \forall z\beta'(x/z)$. It is easy to check that x is not free in $\beta'(x/z)$, x is free for z in $\beta'(x/z)$, and that $\beta'(x/z)(z/x) = \beta'$; hence, the same arguments can be used to prove $\forall z\beta'(x/z) \vdash \forall x\beta'$. Part (i) gives $\forall x\beta' \vdash \forall x\beta$, hence, $\forall z\beta'(x/z) \vdash \forall x\beta$.

If $\alpha = [\beta]_{\bar{x}}$ there are two cases.

(i) $\alpha' = [\beta']_{\bar{x}}$, where β' is a direct variant of β. By induction $\vdash \beta' \to \beta$, so by theorem 4.2.4 $\vdash \forall x_1 \ldots \forall x_n(\beta' \to \beta)$. Hence by axiom P1 and modus ponens we have, $\vdash [\beta' \to \beta]_{\bar{x}} = 1$. Similarly, $\vdash [\beta \to \beta']_{\bar{x}} = 1$. Theorem 4.2.6 yields $\vdash [\beta]_{\bar{x}} = [\beta']_{\bar{x}}$.

(ii) $\alpha' = [\beta'(x_i/z)]_{\bar{x}(x_i/z)}$, where β' is a direct variant of β, z is an object variable not free in β' and free for x_i in β'. By (i), $\vdash [\beta]_{\bar{x}} = [\beta']_{\bar{x}}$, and by axiom (P6) $\vdash [\beta']_{\bar{x}} = [\beta'(x_i/z)]_{\bar{x}(x_i/z)}$. The transitivity of the equality predicate (theorem 4.2.6) gives $\vdash [\beta]_{\bar{x}} = [\beta'(x_i/z)]_{\bar{x}(x_i/z)}$.

Finally, if α' is the result of n direct variations of α we have α' is a direct variant of α_{n-1}, where α_{n-1} is the result of $n-1$ direct variations of α. By induction, if α is a term then $\vdash \alpha = \alpha_{n-1}$, also $\vdash \alpha_{n-1} = \alpha'$. So, $\vdash \alpha = \alpha'$, by the transitivity of the equality predicate. If α is a formula then $\alpha \vdash \alpha_{n-1}$, also $\alpha_{n-1} \vdash \alpha'$. So, $\alpha \vdash \alpha'$. Similarly, $\alpha' \vdash \alpha$. ■

Theorem 4.2.8 *For every formula, α, variable, x, and term, t (of the same type)*

$$\vdash \forall x\alpha \to \alpha(x/t).$$

Proof By definition 4.1.8, $\alpha(x/t) = \alpha'(x/t)$, where α' is a variant of α such that t is free for x in α'. By axiom PC4, $\forall x\alpha' \vdash \alpha(x/t)$, and by theorem 4.2.7, $\forall x\alpha \vdash \forall x\alpha'$. Hence, $\forall x\alpha \vdash \alpha(x/t)$, and by the deduction theorem, $\vdash \forall x\alpha \rightarrow \alpha(x/t)$. ■

4.2.4 Maximal Consistency

To prove the completeness theorem we need the notion of *maximal consistent* sets of formulas. The next few definitions and theorems develop some properties of maximal consistent sets of formulas.

Definition 4.2.9 *A set of* Lp *formulas* Φ *is* inconsistent *if for some* α *both* $\Phi \vdash \alpha$ *and* $\Phi \vdash \neg\alpha$*, otherwise,* Φ *is* consistent. Φ *is* maximal consistent *if* Φ *is consistent and is not a proper subset of any other consistent set of formulas.*

Theorem 4.2.10 *A set of formulas* Φ *is inconsistent iff* $\Phi \vdash \alpha$ *for every formula* α*.*

Proof Standard first order proof holds. ■

Theorem 4.2.11 *For any* Φ *and* α*,*

a) $\Phi, \neg\alpha$ *is inconsistent iff* $\Phi \vdash \alpha$*,*

b) Φ, α *is inconsistent iff* $\Phi \vdash \neg\alpha$*.*

Proof Standard first order proof holds. ■

Theorem 4.2.12 *A set Φ is maximal consistent iff both of the following conditions are satisfied:*

a) *Φ is consistent.*

b) *For every formula α, $\alpha \in \Phi$ or $\neg\alpha \in \Phi$.*

Proof Standard first order proof holds. ■

Theorem 4.2.13 *If Φ is maximal consistent and $\Phi \vdash \alpha$ then $\alpha \in \Phi$.*

Proof Standard first order proof holds. ■

Theorem 4.2.14 *Let Φ be maximal consistent. The following conditions hold:*

1. α and $\neg\alpha$ do not both belong to Φ.
2. If $\neg\neg\alpha \in \Phi$ then $\alpha \in \Phi$.
3. If $\alpha \wedge \beta \in \Phi$ then $\alpha \in \Phi$ and $\beta \in \Phi$.
4. If $\neg(\alpha \wedge \beta) \in \Phi$ then $\neg\alpha \in \Phi$ or $\neg\beta \in \Phi$.
5. If $\forall x \alpha \in \Phi$ then $\alpha(x/t) \in \Phi$ for every term t of the same type as the variable x.
6. For every term t, $(t = t) \in \Phi$.

7. If f is an n-ary function symbol of **Lp** and $t_1, \dots t_{2n}$ are terms of compatible type, then the formula

$$t_1 = t_{n+1} \to \dots \to t_n = t_{2n} \to ft_1 \dots t_n = ft_{n+1} \dots t_{2n}$$

is in Φ.

8. If P is an n-ary predicate symbol and $t_1, \dots, t_{2n}$ are terms of the same type, then the formula

$$t_1 = t_{n+1} \to \dots \to t_n = t_{2n} \to Pt_1 \dots t_n = Pt_{n+1} \dots t_{2n}$$

is in Φ.

Proof By item as follows:

1. Follows immediately from the fact that Φ is consistent.
2. If $\neg\neg\alpha \in \Phi$ we have $\Phi \vdash \alpha$ (as $\neg\neg\alpha \to \alpha$ is a tautology) so by theorem 4.2.13 $\alpha \in \Phi$.
3. If $\alpha \wedge \beta \in \Phi$ we have $\Phi \vdash \alpha$ and $\Phi \vdash \beta$ (again by tautologies) so by theorem 4.2.13 both α and β are in Φ.
4. If $\neg(\alpha \wedge \beta) \in \Phi$ then if $\alpha \in \Phi$ we have $\Phi \vdash \neg\beta$. So either $\neg\alpha \in \Phi$ or $\neg\beta \in \Phi$.
5. If $\forall x\alpha \in \Phi$ then for every term t we have $\Phi \vdash \alpha(x/t)$ by theorem 4.2.8. Thus $\alpha(x/t) \in \Phi$ by theorem 4.2.13.

6-8. Every instance of every axiom of **Lp** (in particular all instances of axioms EQ5, EQ6, and EQ7) is provable, hence, deducible from Φ, thus, in Φ by theorem 4.2.13.

■

4.2.5 Soundness and Completeness of the Proof Theory

Now we have all of the necessary machinery to prove the following existence theorem, from which the completeness theorem follows easily. This is done in a manner similar to the proof in first order logic, i.e., by way of a Henkin construction [3]. Modifications have been made to deal with the definition of the probability function and to handle the two sorted universe.

Theorem 4.2.15 (Existence of a Model) *If Ω is a consistent set of* Lp *formulas then there exists an interpretation σ, with underlying* Lp*-Structure $\mathcal{M}$, which satisfies Ω. That is, $\beta^\sigma = \top$ for all $\beta \in \Omega$.*

Proof First, we extend Ω to a maximal consistent set of formulas Φ which has witnesses, i.e. if $\neg\forall x\alpha \in \Phi$ then for some constant c, $\neg\alpha(x/c) \in \Phi$. To begin it should be noted that **Lp** has only a denumerable number of formulas (as each formula is a string of finite length). We extend **Lp** to a new language **Lp(C)** by adding a denumerable set of new constants $\{c_i | i =$

$0, 1, \ldots\}$. It is clear that Ω is also a consistent set of **Lp(C)** formulas. We fix an ordering $\{\phi_i | i = 0, 1, \ldots\}$ on all the formulas of **Lp(C)**, and define for every n a set Φ_n of **Lp(C)** formulas such that:

1. Φ_i is a subset (may be equal) of Φ_j for all $j > i$.
2. Φ_i is consistent.
3. Only finitely many new constants occur in Φ_i.

First put $\Phi_0 = \Omega$; (1) holds vacuously, (2) holds by assumption and (3) holds because Ω is in **Lp** so has no new constants. At the n-th stage define Φ_{n+1} as follows.

Case 1) If $\Phi_n \cup \{\phi_n\}$ is inconsistent we put $\Phi_{n+1} = \Phi_n$. Clearly (1), (2), and (3) hold for Φ_{n+1}.

Case 2) If $\Phi_n \cup \{\phi_n\}$ is consistent and ϕ_n is not of the form $\neg\forall x\alpha$, make $\Phi_{n+1} = \Phi_n \cup \{\phi_n\}$. Again it is obvious that (1) and (2) hold for Φ_{n+1}. We also have (3) since ϕ_n is a formula of finite length.

Case 3) If $\Phi_n \cup \{\phi_n\}$ is consistent and ϕ_n is of the form $\neg\forall x\alpha$, then by (3) there exists a new constant c which does not occur in Φ_n nor in ϕ_n. With such c we put $\Phi_{n+1} = \Phi_n \cup \{\phi_n, \neg\alpha(x/c)\}$. If x is an object variable then the new constant c is defined to be an o-term; otherwise, x is a field variable and c is defined to be an f-term. It is easy to see that (1) and (3) hold. If Φ_{n+1} were inconsistent then,

> by theorem 4.2.11, we have $\{\Phi_n, \phi_n\} \vdash \alpha(x/c)$; thus, $\{\Phi_n, \phi_n\} \vdash \forall x\alpha$ by axiom PC3 (obviously x is not free in $\alpha(x/c)$). Since $\phi_n = \neg\forall x\alpha$, this contradicts our assumption that $\Phi_n \cup \{\phi_n\}$ is consistent. Thus we see that (2) also holds.

Finally we put $\Phi = \bigcup_{n=1}^{\infty} \Phi_n$. By definition of Φ we have $\Omega = \Phi_0 \subset \Phi$, furthermore, we claim that Φ is consistent. If Φ is inconsistent then we have $\Phi \vdash \alpha$ as well as $\Phi \vdash \neg\alpha$. Since both of these deductions are finite, there must be a finite subset of Φ which is inconsistent (namely the set of formulas which appear in the two deductions), but every finite subset of Φ is contained in some Φ_n and every Φ_n is consistent by (2), contradiction. Not only is Φ consistent it is maximal consistent. To see this let α be any formula of **Lp(C)** not in Φ. For some n, $\alpha = \phi_n$. Since $\alpha \notin \Phi$ it follows that Φ_{n+1} must have been defined as in case 1. Thus $\Phi_n \cup \{\alpha\}$ is inconsistent so $\Phi \cup \{\alpha\}$ must also be inconsistent. This shows that Φ has no proper superset which is consistent.

Since Φ is maximal consistent all of the eight conditions of theorem 4.2.14 hold. Furthermore, if $\neg\forall x\alpha \in \Phi$ then by the construction we have, for some constant c, $\neg\alpha(x/c) \in \Phi$. And, by definition, this constant is a term of the same type as the variable x.

Now we can construct an **Lp**-Structure and an interpretation which satisfies Φ. For each term t we define $[\![t]\!] = \{s | s = t \in \Phi\}$. By theorem 4.2.6 it is deducible that '=' defines an equivalence relation and since Φ, being maximal consistent, is closed under deduction it follows that these are equivalence classes of terms in **Lp(C)**. Since these are equivalence classes, it is clear that $[\![t]\!] = [\![s]\!]$ iff $s = t \in \Phi$.

Lemma 4.2.1 *Let* $t_1, \ldots, t_{2n}$ *be terms of identical type such that* $[\![t_i]\!] = [\![t_{i+n}]\!]$ $(i = 1, \ldots, n)$ *then:*

(a) *For any n-ary function symbol, f, of type compatible with the terms* $t_1, \ldots, t_{2n}$

$$[\![ft_1, \ldots, t_n]\!] = [\![ft_{n+1} \ldots t_{2n}]\!],$$

(b) *For any n-ary predicate symbol, P, of the same type as the terms* $t_1, \ldots, t_{2n}$

$$\textit{if} \quad Pt_1, \ldots, t_n \in \Phi \qquad \textit{then} \quad Pt_{n+1} \ldots t_{2n} \in \Phi.$$

Proof The assumption is that $t_i = t_{i+n} \in \Phi$ $(i = 1, \ldots, n)$, so, using condition (7) of theorem 4.2.14 and applying (1) and (4) n times (along with the definition of $\rightarrow$), we get $(ft_1 \ldots ft_n) = (ft_{n+1} \ldots t_{2n}) \in \Phi$. Thus $[\![ft_1 \ldots ft_n]\!] = [\![ft_{n+1} \ldots t_{2n}]\!]$. Similarly, (b) follows from (8) and applying (1) and (4) $n + 1$ times. ■

Now we define an interpretation function σ and a **Lp**-Structure $\mathcal{M}$ which satisfies Φ. As the set of objects $\mathcal{O}$ we take the set of all equivalence classes of o-terms, i.e.,

$$\mathcal{O} = \Big\{[\![t]\!] \colon t \text{ is an } o\text{-term}\Big\}.$$

As the field of numbers $\mathcal{F}$ we take the set of all equivalence classes of f-terms, i.e.,

$$\mathcal{F} = \Big\{[\![t]\!]: t \text{ is an } f\text{-term}\Big\}.$$

Note, $\mathcal{F}$ and $\mathcal{O}$ include the new constants added by case 3.

If f is an n-ary object function symbol of **Lp** we define an n-ary operation on $\mathcal{O}$ as its interpretation by putting

$$f^\sigma\big([\![t_1]\!], \ldots, [\![t_n]\!]\big) = [\![f t_1 \ldots t_n]\!].$$

If f is a field function symbol or a measuring function symbol then the operation is defined on $\mathcal{F}$ and on $\mathcal{O} \mapsto \mathcal{F}$ respectively. By the lemma this definition is independent of the particular choice of representatives of the equivalence classes $t_1, \ldots, t_n$.

If P is an n-ary object predicate symbol we define an n-ary relation on $\mathcal{O}$ by putting

$$\langle [\![t_1]\!], \ldots, [\![t_n]\!] \rangle \in P^\sigma \quad \text{iff} \quad P t_1 \ldots t_n \in \Phi.$$

Similarly if P is an n-ary field predicate. The lemma shows that this definition is independent of the choice of $t_1, \ldots, t_n$.

For each variable, x, we put $x^\sigma = [\![x]\!]$, where x can be a variable of either type.

Finally we define the sequence of probability functions μ_n on any set of $\mathcal{O}^n$ defined by a formula α by

$$\mu_n\Big\{\langle\llbracket a_1\rrbracket,\ldots,\llbracket a_n\rrbracket\rangle\colon \alpha^{\sigma(\vec{x}/\langle\llbracket a_1\rrbracket,\ldots,\llbracket a_n\rrbracket\rangle)} = \top\Big\} = \llbracket[\alpha]_{\vec{x}}\rrbracket.$$

Lemma 4.2.2 *For any n the probability function μ_n is well defined. That is, if A is a set of tuples in $\mathcal{O}^n$ defined by two different formulas α and β then $\mu_n(A)$ is independent of which formula is used.*

Proof By assumption, $A = \{\langle\llbracket a_1\rrbracket,\ldots,\llbracket a_n\rrbracket\rangle\colon \alpha^{\sigma(\vec{x}/\langle\llbracket a_1\rrbracket,\ldots,\llbracket a_n\rrbracket\rangle)} = \top\}$ and also $A = \{\langle\llbracket b_1\rrbracket,\ldots,\llbracket b_n\rrbracket\rangle\colon \beta^{\sigma(\vec{y}/\langle\llbracket b_1\rrbracket,\ldots,\llbracket b_n\rrbracket\rangle)} = \top\}$. By definition, $\mu_n(A) = \llbracket[\alpha]_{\vec{x}}\rrbracket$ also $\mu_n(A) = \llbracket[\beta]_{\vec{y}}\rrbracket$. The claim of the lemma is that $\llbracket[\alpha]_{\vec{x}}\rrbracket = \llbracket[\beta]_{\vec{y}}\rrbracket$. Let $\vec{z} = \langle z_1,\ldots,z_n\rangle$ be a new set of object variables which do not appear in either α or β. There exists two variants α and β, called α' and β' respectively, formed by substituting all the variables $x_i \in \vec{x}$ in α and all the variables $y_i \in \vec{y}$ in β by the new variables $z_i \in \vec{z}$. By theorem 4.1.7, the sets A' and B' (of tuples of $\mathcal{O}^n$) defined by these variants is the same as the set A. Further, by theorem 4.2.7, it is provable that $[\alpha']_{\vec{z}} = [\alpha]_{\vec{x}}$ also $[\beta']_{\vec{z}} = [\beta]_{\vec{y}}$. So by theorem 4.2.13 we have $\llbracket[\alpha']_{\vec{z}}\rrbracket = \llbracket[\alpha]_{\vec{x}}\rrbracket$ also $\llbracket[\beta']_{\vec{z}}\rrbracket = \llbracket[\beta]_{\vec{y}}\rrbracket$. Hence, the claim can be reduced to proving that $\llbracket[\alpha']_{\vec{z}}\rrbracket = \llbracket[\beta']_{\vec{z}}\rrbracket$.

Since

$$\begin{aligned}\{\langle\llbracket c_1\rrbracket,\ldots,\llbracket c_n\rrbracket\rangle\colon \alpha'^{\sigma(\vec{z}/\langle\llbracket c_1\rrbracket,\ldots,\llbracket c_n\rrbracket\rangle)} = \top\} = \\ \{\langle\llbracket c_1\rrbracket,\ldots,\llbracket c_n\rrbracket\rangle\colon \beta'^{\sigma(\vec{z}/\langle\llbracket c_1\rrbracket,\ldots,\llbracket c_n\rrbracket\rangle)} = \top\}\end{aligned}$$

it must be the case that the formulas $\forall z_i \ldots \forall z_n(\alpha' \rightarrow \beta')$ and $\forall z_i \ldots \forall z_n(\beta' \rightarrow \alpha')$ are in Φ. As Φ is maximal consistent, either these formulas or their negations must be in Φ. If their negations are in Φ it is easy to see, using the witness property of Φ and theorem 4.1.9, that these two sets cannot be equal. Using axiom P1 and theorem 4.2.13, the formulas $[\alpha' \rightarrow \beta']_{\bar{z}} = 1$ and $[\beta' \rightarrow \alpha']_{\bar{z}} = 1$ must be in Φ. Thus, by theorem 4.2.6, $[\alpha']_{\bar{z}} = [\beta']_{\bar{z}} \in \Phi$. Hence, by definition, $[\![[\alpha']_{\bar{z}}]\!] = [\![[\beta']_{\bar{z}}]\!]$. ■

This defines each μ_n on all subsets of $\mathcal{O}^n$ defined by formulas of **Lp**. It should also be clear from the construction of $\mathcal{O}$ that μ_n is also defined on each singleton set of $\mathcal{O}^n$, since the formula $[x_1 = t_1 \wedge \ldots \wedge x_n = t_n]_{\bar{x}}$ defines the singleton set $\{\langle [\![t_1]\!], \ldots, [\![t_n]\!] \rangle\}$. In an **Lp**-Structure each μ_n is defined on a field of subsets of $\mathcal{O}^n$, Π_n. However, theorem 4.1.10 shows that the set of subsets defined by the formulas of **Lp** is itself a field of subsets. Hence, μ_n is already defined over a field of subsets which includes all singleton sets as well as all subsets defined by the formulas of **Lp**. That is, Π_n can be taken to be the field of subsets over which μ_n is already defined.

Lemma 4.2.3 *For each term t*

$$t^\sigma = [\![t]\!].$$

Proof The lemma is proved by induction on the complexity of t. If t is a variable x then $x^\sigma = [\![x]\!]$ by the definition of σ. If $t = [\alpha]_{\bar{x}}$ then

$[\alpha]_x{}^\sigma = [\![[\alpha]_x]\!]$, again by definition. If $t = ft_1 \dots t_n$ then

$$\begin{aligned} t^\sigma &= (ft_1 \dots t_n)^\sigma \\ &= f^\sigma(t_1^\sigma \dots t_n^\sigma) && \text{(By Sem. Defn.)} \\ &= f^\sigma([\![t_1^\sigma]\!] \dots [\![t_n^\sigma]\!]) && \text{(Inductive Hypothesis)} \\ &= [\![ft_1 \dots t_n]\!] && \text{(Def of } f^\sigma\text{)} \\ &= [\![t]\!]. \end{aligned}$$

■

Now we can prove that Φ is in fact satisfied by σ. We prove by induction (on the length of a formula β) that

(a) if $\beta \in \Phi$ then $\beta^\sigma = \top$, and

(b) if $\neg\beta \in \Phi$ then $\beta^\sigma = \bot$ (hence $\neg\beta^\sigma = \top$).

1. $\beta = Pt_1 \dots t_n$, where P is a predicate symbol of either type.

   (a) If $Pt_1 \dots t_n \in \Phi$,

   $$\begin{aligned} &\text{then } \langle [\![t_1]\!], \dots, [\![t_n]\!] \rangle \in P^\sigma, && \text{(by def. of } P^\sigma\text{)} \\ &\langle t_1^\sigma, \dots, t_n^\sigma \rangle \in P^\sigma, && \text{(by lemma 4.2.3)} \\ &(Pt_1 \dots t_n) = \top. && \text{(by Sem. Def.)} \end{aligned}$$

   (b) If $\neg Pt_1 \dots t_n \in \Phi$,

   $$\begin{aligned} &\text{then } Pt_1 \dots t_n \notin \Phi, && \text{(by theorem 4.2.14(1))} \\ &\langle [\![t_1]\!], \dots, [\![t_n]\!] \rangle \notin P^\sigma, && \text{(by def. of } P^\sigma\text{)} \\ &\langle t_1^\sigma, \dots, t_n^\sigma \rangle \notin P^\sigma, && \text{(by lemma 4.2.3)} \\ &(Pt_1 \dots t_n) = \bot. && \text{(by Sem. Def.)} \end{aligned}$$

2. $\beta = (s = t)$, where s and t are terms of the same type.

   (a) If $(s = t) \in \Phi$,

   then $[s] = [t]$,

   $s^\sigma = t^\sigma$, (by lemma 4.2.3)

   $(s = t)^\sigma = \top$. (by Sem. Defn.)

   (b) If $(s \neq t) \in \Phi$

   then $[s] \neq [t]$,

   $s^\sigma \neq t^\sigma$, (by lemma 4.2.3)

   $(s = t)^\sigma = \bot$. (by Sem. Defn.)

3. $\beta = \neg\alpha$.

   (a) If $\neg\alpha \in \Phi$

   then $\alpha^\sigma = \bot$, (by ind. hyp.)

   $(\neg\alpha)^\sigma = \top$. (by Sem. Defn.)

   (b) If $\neg\neg\alpha \in \Phi$

   then $\alpha \in \Phi$, (by thm. 4.2.14(2))

   $\alpha^\sigma = \top$, (by ind. hyp.)

   $(\neg\alpha)^\sigma = \bot$. (by Sem. Defn.)

4. $\beta = \alpha \wedge \delta$.

   (a) If $\alpha \wedge \delta \in \Phi$

   then $\alpha \in \Phi$ and $\delta \in \Phi$, (by thm. 4.2.14(3))

$\alpha^\sigma = \top$ and $\delta^\sigma = \top$, (by ind. hyp.)

$(\alpha \wedge \delta)^\sigma = \top$. (by Sem. Defn.)

(b) If $\neg(\alpha \wedge \delta) \in \Phi$

then $\neg\alpha \in \Phi$ or $\neg\delta \in \Phi$, (by thm. 4.2.14(4))

$\alpha^\sigma = \bot$ or $\delta^\sigma = \bot$, (by ind. hyp.)

$(\alpha \wedge \delta)^\sigma = \bot$. (by Sem. Defn.)

5. $\beta = \forall x\alpha$, where x is a variable of either type.

(a) If $\forall x\alpha \in \Phi$

then $\alpha(x/t) \in \Phi$ for every term t of the same type, (by thm 4.2.14(5))

$\alpha(x/t)^\sigma = \top$ for every t, (by ind. hyp.)

$\alpha^{\sigma(x/t')} = \top$ for every t, where $t' = t^\sigma$, (by thm. 4.1.9)

$\alpha^{\sigma(x/[\![t]\!])} = \top$ for every $u \in \mathcal{O}$ $(u \in \mathcal{F})$, (By def. of $\mathcal{O}$)

$(\forall x\alpha)^\sigma = \top$. (By Sem. Defn.)

(b) If $\neg\forall x\alpha \in \Phi$

then, by the witness property of Φ, $\neg\alpha(x/t) \in \Phi$ for some term t (same type),

$\neg\alpha(x/t)^\sigma = \top$ for some t, (by ind. hyp.)

$\alpha^{\sigma(x/t')} = \bot$, where $t' = t^\sigma$, (by thm.4.1.9)

$\alpha^{\sigma(x/[\![t]\!])} = \bot$, (by lemma 4.2.3)

$\alpha^{\sigma(x/u)} = \bot$ for some $u \in \mathcal{O}$ $(u \in \mathcal{F})$, (By Def. of $\mathcal{O}$)

$(\forall x \alpha)^\sigma = \bot.$ (By Sem. Defn.)

Thus $\beta^\sigma = \top$ for all $\beta \in \Phi$. Since Φ is maximal consistent it contains all instances of all axioms. Thus the structure and interpretation constructed satisfies all of these axioms. In particular, since all of the field axioms are true it is clear that $\mathcal{F}$ has the structure of a field. Further, since all of the probability axioms are true it is the case that the functions μ_n are in fact probability functions.

The sequence of probability functions is a sequence of product measures, since every instance of axiom P7 is true. Let $A = \{\vec{a} | \alpha^{\sigma(\vec{x}/\vec{a})} = \top\} \subset \mathcal{O}^n$ and $B = \{\vec{b} | \beta^{\sigma(\vec{y}/\vec{b})} = \top\} \subset \mathcal{O}^m$ be two sets in the domain of μ_n and μ_m respectively, with $\mu_n(A) = z_1$ and $\mu_m(B) = z_2$. It can be seen that the equivalence class of the probability term $[\alpha \wedge \beta]_{\langle \vec{x}, \vec{y} \rangle}$ is equal to the probability of their Cartesian product. Also, we have $[[\alpha \wedge \beta]_{\vec{x}} = z_1]_{\vec{y}} = z_2$ is true, so must be in Φ. Hence, by axiom P7 the probability of the Cartesian product is greater than or equal to $z_1 \times z_2$. It must be shown that it is in fact equal. This can be done by considering the complement of the Cartesian product. This set is not a product set, but it is equal to the union of two product sets. That is, it is equal to the (disjoint) union of $\neg A \times \mathcal{O}^m$ and $A \times \neg B$. Using P7 again, we see that the complement is greater than equal to $1 - z_1 + z_1 \times (1 - z_2)$, which is $1 - z_1 \times z_2$. The result follows from axiom P3.

Axiom P8 insures that the probability functions satisfy the constraint of invariance under permutations. Similarly axiom P9 forces the functions to satisfy the conglomerability condition. This can be seen by examining the semantic interpretation of axiom P9.

Hence, the structure constructed is a valid **Lp**-Structure.

Since Ω is contained in Φ, it is obvious that σ satisfies Ω. That is, $\alpha^\sigma = \top$ for all $\alpha \in \Omega$ as claimed. ∎

The proof theory is both sound and complete. This means that deductions are semantically valid and that deductions exist for all semantic entailments.

Theorem 4.2.16 (Completeness) *If* $\Phi \models \alpha$, *then* $\Phi \vdash \alpha$.

Proof If $\Phi \models \alpha$ then no interpretation satisfies $\{\Phi, \neg\alpha\}$. Hence, by the Existence Theorem, $\{\Phi, \neg\alpha\}$ is inconsistent. Thus, by theorem 4.2.11, $\Phi \vdash \alpha$. ∎

Theorem 4.2.17 (Soundness) *If* $\Phi \vdash \alpha$, *then* $\Phi \models \alpha$.

Proof Let $\phi_1, \ldots, \phi_n$ be a deduction of α from Φ, i.e. $\phi_n = \alpha$. We show by induction on $k = 1, \ldots n$ that $\Phi \models \phi_k$. If ϕ_k is an axiom then we claim that ϕ_k is satisfied by every interpretation. Thus, $\Phi \models \phi_k$. If $\phi_k \in \Phi$ then it is clear that $\Phi \models \phi_k$. The last case is if for some $i, j < k$ we have $\phi_j = \phi_i \rightarrow \phi_k$. By induction $\Phi \models \phi_i$ and $\Phi \models \phi_j$, so, from the semantic definition and the definition of '→', it follows that $\Phi \models \phi_k$. Now, all that remains is to prove the claim that the axioms of **Lp** are satisfied by every interpretation. The first order axioms pose no problem, since **Lp** is an extension of first order logic. The standard proof of the soundness theorem for first order logic suffices to show that these axioms are valid (satisfied by every interpretation). Since in the **Lp**-Structure $\mathcal{F}$ is defined to be an ordered field, it is clear that all of the field axioms are valid. Finally, since each μ_n is defined to be a probability function in the **Lp**-Structure, we can use the semantic definition of the probability terms $[\alpha]_{\mathcal{I}}$ to see that axioms P1-P5 are valid. Theorem 4.1.7 shows that axiom P6 is valid. The fact that the sequence of probability functions is a sequence of product measures yields the validity of axiom P7. The additional constraints (2) and (3) ensure that axioms P8 and P9 are valid. ■

4.3 Properties of the Probability Terms

This section presents some simple lemmas which demonstrate some properties of the probability terms. These results will be used in the examples which follow. The existence of a completeness proof allows a proof of these lemmas from the semantics; the corresponding syntactic proof is guaranteed to exist. In these cases, a proof from the semantics is much simpler, as it just requires using some notions from set theory and probability theory, whereas, a syntactic proof would involve a lot of symbolic manipulation, a task more suited to an automatic theorem prover.

Lemma 4.3.1 *The following are provable in* **Lp**.

a) $[\alpha]_{\vec{x}} \leq 1$.

b) $[\alpha \wedge \beta]_{\vec{x}} \leq [\alpha]_{\vec{x}}$ and $[\alpha \wedge \beta]_{\vec{x}} \leq [\beta]_{\vec{x}}$.

c) $[\alpha \vee \beta]_{\vec{x}} \geq [\alpha]_{\vec{x}}$ and $[\alpha \vee \beta]_{\vec{x}} \geq [\beta]_{\vec{x}}$.

d) $[\alpha \vee \beta]_{\vec{x}} = [\alpha]_{\vec{x}} + [\beta]_{\vec{x}} - [\alpha \wedge \beta]_{\vec{x}}$.

Proof All of these results can be simply deduced from the fact that semantically the probability terms represent assignments of probability. That is, each probability term represents the probability of a corresponding set of objects in $\mathcal{O}^n$. Hence, all of these results follow from the properties of the probability functions μ_n, (their non-standard features do not affect these results). Equivalently they can be deduced from the probability and field axioms, in a manner similar to the proof of theorem 4.2.6. ■

That these results are provable in **Lp** is an important point. They indicate that the probability functions have many of the familar properties of ordinary real valued probability functions, even though they assume values in a field which is only defined by abstract field axioms, a field which is not necessarily the field of real numbers. This means that when numeric values are available, a practical reasoning system could use the arithmetic hardware already built into computers. Any new numeric probability computed by arithmetic calculation from other numeric probabilities, using the familiar properties of probability functions, will be a valid deduction. This deduction will be inferred much faster than if it was inferred through symbolic manipulation of the field axioms. The advantage of having the field axioms, besides the contribution of the field to representational power, arises from the fact that in many, if not most, situations numeric probabilities are not available. In this case the field axioms allow one to reason with whatever information is available. For example, if the knowledge base contained the set of statements $\{[P(x)]_x > [Q(x)]_x, [Q(x)]_x > [R(x)]_x\}$, then it would be possible, using axiom F11, to infer $[P(x)]_x > [R(x)]_x$, even though no numeric values were available.

Theorem 4.3.1 (Bayes' Theorem) *Using definition 3.2.2, the following is provable in* **Lp***:*

$$([\alpha]_{\vec{x}} \neq 0 \wedge [\beta]_{\vec{x}} \neq 0) \rightarrow [\beta|\alpha]_{\vec{x}} = [\alpha|\beta]_{\vec{x}} \times \frac{[\beta]_{\vec{x}}}{[\alpha]_{\vec{x}}}.$$

This theorem shows that the powerful mechanisms of Bayesian inference are also valid in **Lp**. Bayesian analysis is useful when numeric probabilities are available. It requires a certain minimum amount of probabilistic information (although, as Pearl has shown [55], the information requirements can be made reasonable if knowledge of dependencies are also available).

Inference engines formally based on Bayes' theorem and the laws of probability can be used on numeric probabilities expressed in **Lp**. Since both the probability axioms and Bayes' theorem are valid in **Lp**, the conclusions obtained from such inference engines will be valid deductions in **Lp**.

The next lemma shows that when $\beta \rightarrow \lambda$, λ does not affect the conditional probability.

Lemma 4.3.2 *If* $\beta \vdash \lambda$ *then* $[\alpha|\beta \wedge \lambda]_{\vec{x}} = [\alpha|\beta]_{\vec{x}}$, *given that* $[\beta]_{\vec{x}} \neq 0$.

Proof: Since $[\beta]_{\vec{x}} > 0$, $\{\vec{a}|(\beta)^{\sigma(\vec{x}/\vec{a})} = \top\}$ is not empty. Let $\vec{c}$ be a member of this set. By the soundness theorem $\beta \models \lambda$, i.e., if $\beta^{\tau} = \top$ then $\lambda^{\tau} = \top$ for any interpretation τ. Hence, we have that $\lambda^{\sigma(\vec{x}/\vec{c})} = \top$, and, by the semantic definition, $\vec{c} \in \{\vec{a}|(\beta \wedge \lambda)^{\sigma(\vec{x}/\vec{a})} = \top\}$. Therefore, we have $\{\vec{a}|(\beta)^{\sigma(\vec{x}/\vec{a})} = \top\} \subset \{\vec{a}|(\beta \wedge \lambda)^{\sigma(\vec{x}/\vec{a})} = \top\}$. Clearly, the opposite containment also holds, hence, the two sets are equal. By the semantic definition we have $[\beta]_{\vec{x}} = [\beta \wedge \lambda]_{\vec{x}}$, and it is easy to show that $[\beta \wedge \alpha]_{\vec{x}} = [\beta \wedge \lambda \wedge \alpha]_{\vec{x}}$ also. The lemma follows from the definition of conditional probabilities. ■

The last lemma shows that deductive consequences always have greater conditional probability.

Lemma 4.3.3 *If* $\vdash \forall x_1 \ldots x_n(\beta \rightarrow \lambda)$ *then* $[\lambda|\alpha]_{\vec{x}} \geq [\beta|\alpha]_{\vec{x}}$.

Proof: Using the soundness theorem it is easy to show that $\{\vec{a} | (\beta \wedge \alpha)^{\sigma(\vec{x}/\vec{a})} = \top\}$ is a subset of $\{\vec{a} | (\lambda \wedge \alpha)^{\sigma(\vec{x}/\vec{a})} = \top\}$. Since μ_n is a probability function $[\lambda \wedge \alpha]_{\vec{x}} \geq [\beta \wedge \alpha]_{\vec{x}}$, and the result follows from the definition of conditionals. ∎

4.4 Examples of Reasoning with the Statistical Knowledge

Example 4.1 *Nilsson's Probabilistic Entailment.*

Nilsson [53] develops a probability logic based on the possible worlds approach. He shows how the probabilities of sentences in the logic are constrained by known probabilities, i.e., constrained by the probabilities of a base set of sentences. For example, if $[P \wedge Q] = 0.5$, then the values of $[P]$ and $[Q]$ are both constrained to be ≥ 0.5. Nilsson demonstrates how the implied constraints of a base set of sentences can be represented in a canonical manner, as a set of linear equations. These linear equations can be used to identify the strongest constraints on the probability of a new sentence, i.e., the tightest bounds on its probability. These constraints are, in Nilsson's terms, probabilistic entailments.

Nilsson gives some approximate methods for calculating these entailments, as well as noting that the methods of linear programming can give exact solutions. The important point, however, is that these bounds are simply consequences of the laws of probability. In fact, the theorem

$$[\alpha \vee \beta]_{\vec{x}} = [\alpha]_{\vec{x}} + [\beta]_{\vec{x}} - [\alpha \wedge \beta]_{\vec{x}},$$

along with the fact that probabilities are not negative, gives the full set of constraints from which all probabilistic entailments are derived. This theorem is true in **Lp** (lemma 4.3.1). And, since the proof theory of **Lp** is complete, probabilistic entailments can be deduced in **Lp**. Numerically the constraints are identical, i.e., the best bounds deducible in **Lp** are same numbers as the best probabilistic entailments.

For example, if the base set in Nilsson's logic is $\{[P]{=}0.6, [P \rightarrow Q]{=}0.8\}$, probabilistic entailment gives the conclusion $0.4 \leq [Q] \leq 0.8$. If we write the symbols P and Q as one place predicates, then in **Lp** the knowledge could be represented by the following set: $\{[P(x)]_x = 0.6, [P(x) \rightarrow Q(x)]_x = 0.8\}$.

From this knowledge it is easy to deduce the bounds $[0.4, 0.8]$ on the probability term $[Q(x)]_x$.[3]

Example 4.2 *Simple reasoning with empirical generalizations (defaults).*

---

[3] It should be noted that Nilsson's probabilities are subjective while the probabilities in **Lp** are empirical, hence they are not quite comparable. However, the next chapter will demonstrate a mechanism of generating 'subjective' probabilities from the empirical probabilities encoded in **Lp**. It will be shown how reasoning with these 'subjective' probabilities can be performed by reasoning with the base empirical knowledge encoded in **Lp**. Hence, this formalism is in fact capable of duplicating all of the reasoning possible with Nilsson's system.

1. If the statement "*P*'s are typically *Q*'s" is given the statistical interpretation that more than **c**% of all *P*'s are also *Q*'s, where **c** is some number close to 1, then the opposite conclusion, that "*P*'s are typically not *Q*'s," can be proved to be false.[4] That is,

$$[Q(x)|P(x)]_x > \mathbf{c} \vdash \neg\Big([\neg Q(x)|P(x)]_x > \mathbf{c}\Big).$$

The derivation follows from axiom P3.

2. Similarly, the if the statement "*P*'s are *Q*'s" is asserted then the statement "*P*'s are typically not *Q*'s," can be proved to be false. For example, "penguins are birds" implies that "penguins are typically not birds" is false.

$$\Big(\forall x(penguin(x) \rightarrow bird(x)\Big) \vdash \neg\Big([bird(x)|penguin(x)]_x > \mathbf{c}\Big).$$

The derivation follows from axioms P1 and P3.

Example 4.3 *More complex reasoning with generalizations.*

1. The knowledge, "most ravens are black" along with "black objects are not white," can be used to deduce that "most ravens are not white."

$$\{[black(x)|raven(x)]_x > \mathbf{c},$$

---

[4] The fact that most non-monotonic formalisms allow both of these statements to be asserted without contradiction has been noted, and cited as a weakness, by both Touretzky et al. [78] and Delgrande [16].

$$\forall x(black(x) \rightarrow \neg white(x))\}$$

$$\vdash [\neg white(x)|raven(x)]_x > \mathbf{c}.$$

This can be shown with an argument similar to lemma 4.3.3.

2. The knowledge, "most birds fly" along with "penguins do not fly", can be used to deduce that "most birds are not penguins."

$$\{[fly(x)|bird(x)]_x > \mathbf{c},$$

$$\forall x(penguin(x) \rightarrow \neg fly(x))\}$$

$$\vdash [\neg penguin(x)|bird(x)]_x > \mathbf{c}.$$

Example 4.4 *Weighing of evidence with explicit assumptions of independence.*

Let $F(x)$ represent the assertion that x is a car with faulty hydraulics, $Sq(x)$ the assertion that x is a car with squeaky brakes, and $Sp(x)$ the assertion that x is a car with spongy brakes. Given knowledge about the prior probabilities of faulty hydraulics, squeaky brakes, and spongy brakes, the conditional probabilities of squeaky brakes and spongy brakes given faulty hydraulics, as well as the knowledge that the probability of spongy brakes is independent of squeaky brakes, both unconditionally and when given that the car has faulty hydraulics, then the probability that a car has faulty hydraulics when it is observed to have both squeaky and spongy brakes can be deduced using Bayes theorem.

$$
\begin{aligned}
&\{[Sq(x)|F(x)]_x = \mathbf{Cd}^{Sq},\\
&\quad [Sp(x)|F(x)]_x = \mathbf{Cd}^{Sp},\\
&\quad [F(x)]_x = \mathbf{c}^F, [Sq(x)]_x = \mathbf{c}^{Sq},\\
&\quad [Sp(x)]_x = \mathbf{c}^{Sp},\\
&\quad [Sq(x) \wedge Sp(x)]_x = [Sq(x)]_x \times [Sp(x)]_x,\\
&\quad [Sq(x) \wedge Sp(x)|F(x)]_x = [Sq(x)|F(x)]_x \times [Sp(x)|F(x)]]_x\}\\
&\vdash [F(x)|Sq(x) \wedge Sp(x)]_x = \frac{\mathbf{Cd}^{Sq} \times \mathbf{Cd}^{Sp} \times \mathbf{c}^F}{\mathbf{c}^{Sq} \times \mathbf{c}^{Sp}}
\end{aligned}
$$

This example is, of course, very simple and requires a lot of statistical knowledge, however, it serves to illustrate the point that the techniques of Bayesian analysis are subsumed by the deductive proof theory.

Chapter 5

Belief Formation

As mentioned in chapter 2, **Lp** cannot express an assignment of probability to a closed formula, e.g., a probability assignment to the formula $Bark(Fido)$. The probability terms state the probability of the set of objects for which a formula is true. In their semantic definition (section 3.5), there is no mention of which individuals satisfy the formula. These probability terms express empirical probabilities over sets of individuals; such probabilities do not apply to particular individuals.[1] This limitation of empirical probabilities has long been noted, by various writers who have adopted an empirical interpretation of probabilities (see Kyburg [37, page 8]). In fact, the semantics of **Lp** allows a formal demonstration of this limitation, as is shown by the following lemma.

---

[1] This can be contrasted with universal quantification. A universally quantified formula is true for all individuals; so, it is necessarily true for any particular individual.

Lemma 5.0.1 *If α is a closed formula then $[\alpha]_{\vec{x}} = 0$ or 1.*

Proof By the semantic definition, for any interpretation σ:

$$([\alpha]_{\vec{x}})^{\sigma} = \mu_n\{\vec{a} | \alpha^{\sigma(\vec{x}/\vec{a})} = \top\}.$$

Since α has no free variables, σ and $\sigma(\vec{x}/\vec{a})$ will agree on all the free variables of α, for any $\vec{a}$. Hence, by theorem 4.1.2, $\alpha^{\sigma} = \alpha^{\sigma(\vec{x}/\vec{a})}$. Either $\alpha^{\sigma} = \top$ or $\alpha^{\sigma} = \bot$, since σ is an interpretation and α is a formula. Thus, the above set of $\vec{a}$ is either all of $\mathcal{O}^n$ or the empty set, and, for any μ_n, the probability is either 0 or 1. ■

Even though probabilities cannot be assigned to closed formulas in **Lp**, there is clearly a need for such probability assignments. For example, the assertion $Bark(Fido)$ may not be deducible from the knowledge base, but it may be necessary to assign it some degree of belief, e.g., as a guide to action. This assertion can be assigned a reasonable degree of belief by using the statistical knowledge that most dogs bark (given that *Fido* is a dog). This chapter presents a general inductive mechanism of belief formation, which can use the non-specific statistical information expressed in **Lp** to generate degrees of belief in closed formulas (sentences) which cite specific individuals.

When the sentence is not deducible from the knowledge base, the mechanism is capable of generating degrees of belief in the range 0–1, or other information, e.g., interval information or comparative information. In fact, the information about the degree of belief can be any information about a number expressible in **Lp**. The extreme degrees of belief, i.e., 0 or 1, can be generated when the sentence, or its negation, is deducible. That is, when the sentence is deducible the mechanism can generate a degree of belief representative of the entailed truth value.

First, the mechanism itself is presented. Then, it is demonstrated how it can be justified by the semantics of **Lp**. The last section gives some examples which illustrate the generality of the formalism.

5.1 Belief Formation

First, we define a belief function, $\mathcal{B}$, which maps pairs of **Lp** sentences to numbers in the closed interval $[0,1]$.

Definition 5.1.1 (Belief Function) *Let $\mathcal{B}(\alpha|\beta)$ denote the degree of belief in the sentence α given the base knowledge β (also a sentence). This degree is a number in the closed interval* $[0,1]$.

It will be seen, however, that the result of the belief function is dependent not only on its two arguments, α and β, but also on a background knowledge base.

5.1.1 Inductive Evaluation Function

It is convenient to use a specialization of the notation developed for substitution in section 4.1. If $\vec{c} = \langle c_1, \ldots, c_n \rangle$ is a vector of distinct object constants and $\vec{x}$ a vector of distinct object variables, then denote by $\alpha(\vec{c}/\vec{x})$ the new formula which results from substituting x_i for every occurrence of c_i in the formula α. (If the c_i's were considered to be free variables in α, definition 4.1.8 would yield the same formula.) Also, let KB denote the set of closed **Lp** formulas which comprise the knowledge base.

The belief function is evaluated through the following inductive principle.

Definition 5.1.2 (Inductive Principle) *Given that a closed formula, α, contains the vector of object constants $\vec{c}$ (and no other object constants), the degree of belief $\mathcal{B}(\alpha|\beta)$ is assigned a value equal to the following* Lp *probability term:*

$$\mathcal{B}(\alpha|\beta) = [\alpha(\vec{c}/\vec{x})|\beta(\vec{c}/\vec{x})]_{\vec{x}},$$

where $\vec{x}$ is a vector of object variables which do not occur in α or β.

This inductive principle has a simple intuitive interpretation. The degree of belief in $\alpha(\vec{c})$ given the knowledge $\beta(\vec{c})$, $\mathcal{B}(\alpha|\beta)$, is equal to the probability that a random tuple $\vec{x}$, with all the properties β given for $\vec{c}$, will have properties α.

The degree of belief is generated through an inductive assumption of randomization. That is, the particular tuple of individuals mentioned in the formula α, $\vec{c}$, is considered to be a randomly selected tuple from the set of tuples $\{\vec{a}|(\beta(\vec{c}/\vec{x})^{\sigma(\vec{x}/\vec{a})} = \top\}$. For example, if we had the statistical knowledge $[Fly(x)|Bird(x)]_x > .75$, the inductive principle would assign the degree of belief $\mathcal{B}(Fly(Tweety)|Bird(Tweety))$ a value $> .75$. This value, is based on the statistical knowledge that if a bird was selected at random (i.e., an individual selected at random from the set $\{a|Bird(x)^{\sigma(x/a)} = \top\}$) then there is a $>$ 75% chance that it would be able to fly. The inductive assumption is that it is reasonable to use this value as the degree of belief in the assertion that this particular bird, *Tweety*, can fly, since the only knowledge being used about *Tweety* is that he is a bird. This example also demonstrates that the belief function depends not only on the two sentences which are its arguments, but also on background statistical knowledge which determine the value of the probability term that is the assigned degree of belief.

If an agent is interested in a particular sentence α, it would seem that there is an impossibly large set of different degrees of belief, $\mathcal{B}(\alpha|\beta)$, which could be formed about α, each one based on a different sentence β. However, the knowledge base will not contain any useful information about the values of most of these degrees of belief, i.e., all that will be deducible about the probability terms generated by the inductive principle is that they are in the closed interval 0–1. Furthermore, the following lemma shows that it is only knowledge about the particular set of individuals appearing in α that is relevant.

Lemma 5.1.1 *If no $x_i \in \vec{x}$ is free in λ then $[\alpha|\beta \wedge \lambda]_{\vec{x}} = [\alpha|\beta]_{\vec{x}}$, assuming that $[\beta \wedge \lambda]_{\vec{x}} \neq 0$.*

Proof: Since $[\beta \wedge \lambda]_{\vec{x}} > 0$, $\{\vec{a}|(\beta \wedge \lambda)^{\sigma(\vec{x}/\vec{a})} = \top\}$ is not empty. Let $\vec{a}'$ be a member of this set; since $(\beta \wedge \lambda)^{\sigma(\vec{x}/\vec{a}')} = \top$, we have by the semantic definition $\lambda^{\sigma(\vec{x}/\vec{a}')} = \top$. Since no member of $\vec{x}$ is free in λ, $\lambda^{\sigma(\vec{x}/\vec{u})} = \top$ for all $\vec{u} \in \mathcal{O}^n$ by theorem 4.1.2. Therefore, $\vec{c} \in \{\vec{a}|(\beta \wedge \lambda)^{\sigma(\vec{x}/\vec{a})} = \top\}$ iff $\vec{c} \in \{\vec{a}|\beta^{\sigma(\vec{x}/\vec{a})} = \top\}$. It is clear from this result that $\{\vec{a}|(\alpha \wedge \beta \wedge \lambda)^{\sigma(\vec{x}/\vec{a})} = \top\} = \{\vec{a}|(\alpha \wedge \beta)^{\sigma(\vec{x}/\vec{a})} = \top\}$. Thus, by definition 3.2.2, $[\alpha|\beta \wedge \lambda]_{\vec{x}} = [\alpha|\beta]_{\vec{x}}$. ■

This lemma implies that when inducing a degree of belief in a sentence α which contains the vector of object constants $\vec{c}$, only those sentences which contain one or more of the constants c_i need be used as base knowledge. For example, if the knowledge base is

$$
\begin{aligned}
\{ \quad & [P(x)|Q(x)]_x = 0.9, \\
& [P(x)|R(x)]_x = 0.5, \\
& [P(x)|Q(x) \wedge R(x)]_x = 0.8, \\
& Q(Tim), R(Tim),
\end{aligned}
$$

$$Q(John), \neg R(John) \quad \},$$

then when inducing a degree of belief in the sentence $P(Tim)$, it is only the knowledge $Q(Tim)$ and $R(Tim)$ which need be used. Let Γ represent the conjunction of all the sentences in the knowledge base, then $\mathcal{B}(P(Tim)|\Gamma)$ has a value equal to the probability term

$$\begin{aligned}[P(z) \quad | \quad & [P(x)|Q(x)]_x = 0.9 \wedge \\ & [P(x)|R(x)]_x = 0.5 \wedge \\ & [P(x)|Q(x) \wedge R(x)]_x = 0.8 \wedge \\ & Q(John) \wedge \neg R(John) \\ & Q(z) \wedge R(z)]_z\end{aligned}$$

By the lemma, this is equal to $[P(z)|Q(z) \wedge R(z)]_z$. By axiom P6, $[P(z)|Q(z) \wedge R(z)]_z = [P(x)|Q(x) \wedge R(x)]_x$, which from the information in the knowledge base is equal to 0.8. Hence, $\mathcal{B}(P(Tim)|\Gamma)$, which is based on all of the knowledge in the knowledge base is equal to $\mathcal{B}(P(Tim)|Q(Tim) \wedge R(Tim))$, which is based only on those sentences which contain the constant *Tim*.

5.1.2 Preference Criterion

Even with the result of the previous lemma, the inductive principle can still yield many different degrees of belief in the sentence α. That is, in general, the knowledge base will contain information about various probability terms $[\alpha(\vec{c}/\vec{x})|\beta(\vec{c}/\vec{x})]_{\vec{x}}$, where β includes one or more of the constants c_i. In some cases these degrees of belief may be conflicting or contradictory. For example, the knowledge base may be the set

$$\begin{aligned} \{ \quad & [Bark(x)|Dingo(x)]_x < 0.1, \\ & [Bark(x)|Dog(x)]_x > 0.9, \\ & \forall x\, Dingo(x) \rightarrow Dog(x), \\ & Dingo(Fido) \quad \}. \end{aligned}$$

In this case, both the sentences $Dingo(Fido)$ and $Dog(Fido)$, deducible from the knowledge base, contain the constant $Fido$. Furthermore, the knowledge base contains non-trivial information about the value of the belief function evaluated on both sentences. This information is contradictory. One sentence yields a low level of belief in the assertion $Bark(Fido)$ while the other yields a high level of belief. In some situations it may be impossible to choose between these competing degrees of belief.

The intuitive interpretation of $\mathcal{B}(\alpha|\beta)$, however, yields a natural preference criterion which in many cases can decide which degree of belief is better. Intuitively $\mathcal{B}(\alpha|\beta)$ represents the degree of belief in α given the knowledge β. Hence, it is reasonable to prefer degrees of belief based on more knowledge, which yields the following criterion:

Definition 5.1.3 (Preference Criterion) *The degree of belief* $\mathcal{B}(\alpha|\beta)$ *is* *to be preferred to the degree of belief* $\mathcal{B}(\alpha|\delta)$, *written*

$$\mathcal{B}(\alpha|\beta) >> \mathcal{B}(\alpha|\delta), \quad \textit{if } KB \vdash \forall\vec{x}\beta(\vec{c}/\vec{x}) \rightarrow \delta(\vec{c}/\vec{x}).$$

In the previous example it can be seen that $KB \vdash \forall x(Dingo(x) \rightarrow Dog(x))$. Hence,

$$\mathcal{B}(Bark(Fido)|Dingo(Fido)) >> \mathcal{B}(Bark(Fido)|Dog(Fido)),$$

and the preferred degree of belief has a low value (< 0.1).

So far we have not imposed any restrictions on the sentences, β, which can act as base knowledge. Lemma 5.1.1 shows that some sentences are irrelevant, and the preference criterion asserts that some sentences are to be preferred, but neither impose any restrictions on the set of sentences which can be used as base knowledge. However, to attain a coherent mechanism of belief formation a simple restriction is needed.

The restriction is that the degree of belief be well-founded.

Definition 5.1.4 *A degree of belief* $\mathcal{B}(\alpha|\beta)$ *is* well-founded *if* $KB \vdash \beta$.

This restriction eliminates the possibility of forming beliefs based on conjecture.[2] The obvious contradiction which could occur if this was allowed is that when inducing a degree of belief in α, α could itself be used as the base knowledge. $\mathcal{B}(\alpha|\alpha)$ always has degree one, but it is seldom informative. The only situation in which it is informative is when it is well-founded.

---

[2] Forming beliefs using conjectures, i.e., sentences which are not deducible from the knowledge base, may have some uses. In particular, such non-well-founded beliefs could be useful for hypothetical reasoning.

For any belief $\mathcal{B}(\alpha|\beta)$, the belief $\mathcal{B}(\alpha|\beta \wedge \alpha)$ is to be preferred. If $KB \vdash \alpha \wedge \beta$ then $\mathcal{B}(\alpha|\beta \wedge \alpha)$ is well-founded and its value is 1. Similarly if $KB \vdash \neg\alpha \wedge \beta$, then $\mathcal{B}(\alpha|\beta \wedge \neg\alpha)$ is preferred, well-founded, and has value 0. That is, when the truth value of α is entailed by the knowledge base the most preferred, well-founded degree of belief about α is representative of that truth value. In this case, degrees of belief about α based on the same sentence α are informative.

5.1.3 Properties of the Mechanism

The entire mechanism of belief formation can be viewed as a process over time, with deduction in **Lp** playing the most important role. The process would involve the continuing deduction in **Lp** of new base knowledge, β, about the individuals, $\vec{c}$, which appear in the sentence of interest, α. Interleaved with the deduction of new base knowledge would be the deduction of information about the degrees of belief generated by the new base knowledge. The inductive principle assigns as the value of each new degree of belief a particular probability term. These probability terms are terms of **Lp**, hence, **Lp** deduction can be used to generate information about them.

A key factor in this process of belief formation is the organization of the knowledge base. Work by Schubert and his associates (see, e.g., [15]) has demonstrated that an efficient organization of the knowledge base allows very rapid deduction of certain types of information, independent of the size of the knowledge base. The next chapter will demonstrate one such organizational scheme, an inheritance net, which allows, through the above process, rapid deduction of information about the degrees of belief in a simple class of assertions.

The preference criterion allows the mechanism of belief formation to behave non-monotonically. If the mechanism is inducing degrees of belief in a sentence α, those degrees can change radically with the addition of new information to the knowledge base. New information allows the formation of new well-founded degrees of belief. These new degrees of belief may have values, assigned to them by the inductive principle, which are very different from any of the previously available degrees of belief, thus calling into question these previously held levels of belief. The new degrees of belief may even be preferred over the previously available degrees, thus superceding all of the previously held levels of belief. For example, if the sentence α is added to the knowledge base, all the old degrees of belief, $\mathcal{B}(\alpha|\beta)$, would be superceded by new preferred degrees of belief which conjoin the sentence α to the base knowledge: $\mathcal{B}(\alpha|\beta \wedge \alpha)$. The value of all of these new preferred degrees of belief is one.

5.2 Semantic Justification

The two parts of the inductive mechanism, the inductive principle, and the preference criterion can both be given a justification based on the semantics of **Lp**.

The inductive principle can be justified by reference to long term behaviour. $\mathcal{B}(\alpha|\beta)$ is assigned a value by considering $\vec{c}$, the vector of individuals which appear in α, to be a random member of the set of vectors which satisfy $\beta(\vec{c}/\vec{x})$. It is easy to show, using the laws of large numbers (see, e.g., Chung [10, Theorem 5.4.2]), that if vectors of objects are drawn at random from the set of object vectors $\mathcal{O}^n$ ($n =$ the size of $\vec{c}$), then the proportion of vectors which satisfy the formula $\alpha(\vec{c}/\vec{x}) \wedge \beta(\vec{c}/\vec{x})$ to the vectors which just satisfy $\beta(\vec{c}/\vec{x})$ approaches in the limit $[\alpha(\vec{c}/\vec{x})|\beta(\vec{c}/\vec{x})]_{\vec{x}}$. That is, if we just count those vectors which are in $\beta(\vec{c}/\vec{x})$, then the proportion of them which satisfy $\alpha(\vec{c}/\vec{x})$ approaches in the limit the value assigned as the degree of belief.

The preference criterion can be given a simple justification, which can be viewed either syntactically or semantically. Syntactically, if $\beta \rightarrow \delta$ then $\beta \leftrightarrow \beta \wedge \delta$. Thus, $\mathcal{B}(\alpha|\beta)$ is equal to $\mathcal{B}(\alpha|\beta \wedge \delta)$ (see lemma 4.3.2). That is, the degree of belief in α based only on the knowledge β is equivalent to a degree of belief based on both β and δ; whereas, $\mathcal{B}(\alpha|\delta)$ is the degree of belief in α based only on δ. In this sense, $\mathcal{B}(\alpha|\beta)$ is based on more knowledge, and thus, should be a preferred degree of belief.

Semantically, when the degree of belief is assigned a value, we are considering the constants which appear in α to be indistinguishable from all of the vectors which satisfy $\beta(\vec{c}/\vec{x})$. If $\forall\vec{x}\beta(\vec{c}/\vec{x}) \rightarrow \delta(\vec{c}/\vec{x})$ then it is the case that the set of vectors satisfying $\beta(\vec{c}/\vec{x})$ is included in the set of vectors satisfying $\delta(\vec{c}/\vec{x})$. Hence, we are losing less information when $\vec{c}$ is considered to be indistinguishable from the vectors which satisfy $\beta(\vec{c}/\vec{x})$ than when $\vec{c}$ is considered to be indistinguishable from the vectors which satisfy $\delta(\vec{c}/\vec{x})$, since, in the former case, we are randomizing over a subset of $\delta(\vec{c}/\vec{x})$.

5.3 Belief Formation—Examples

Example 5.1 *Classical Bayesian Analysis.*

For simplicity only two hypotheses are dealt with. Let $H_1(x)$ and $H_2(x)$ represent an exhaustive and mutually exclusive set of hypotheses which explain some evidence $E(x)$. This knowledge can be represented in **Lp** by the set

$$\{[H_1(x) \vee H_2(x) | E(x)]_x = 1, [H_1(x) \wedge H_2(x)]_x = 0\}.$$

The following derivation is deducible from this knowledge in **Lp** (i.e., a deduction can be constructed for each step in this derivation):

$$[E(x)]_x = [E(x) \wedge H_1(x) \vee H_2(x)]_x$$

$$\begin{aligned} &= [E(x) \wedge H_1(x) \vee E(x) \wedge H_2(x)]_x \\ &= [E(x) \wedge H_1(x)]_x + [E(x) \wedge H_2(x)]_x \\ &= [E(x)|H_1(x)]_x [H_1(x)]_x + [E(x)|H_2(x)]_x [H_2(x)]_x \end{aligned}$$

Hence, the generalized Bayes' rule for the probability of causes is provable in **Lp**, i.e.,

$$[H_i(x)|E(x)]_x = \frac{[E(x) \wedge H_i(x)]_x}{[E(x)]_x} = \frac{[E(x)|H_i(x)]_x [H_i(x)]_x}{\sum_i [E(x)|H_i(x)]_x [H_i(x)]}.$$

If the values of these probability terms are known then degrees of belief of the form $\mathcal{B}(H_i(c)|E(c))$, where c is any object constant, can be evaluated. That is, if we know that c has property E then we can deduce a level of belief in the assertion that it also has property H_i.

Example 5.2 *Comparative Probabilities.*

If our knowledge base consisted of a set of rankings, e.g., the set

$$\{[H_1(x)|E(x)]_x > [H_2(x)|E(x)]_x, [H_2(x)|E(x)]_x > [H_3(x)|E(x)]_x\},$$

then using the field axioms, it is possible to rank degrees of belief of the form $\mathcal{B}(H_i(c)|E(c))$. That is, given $E(c)$ for some constant c, it is deducible that the degree of belief in, e.g., $H_1(c)$ is greater than the degree of belief in $H_3(c)$. Rankings of this sort may be sufficient when all that is required is to choose among the alternative hypotheses, for example, when choosing between competing diagnoses.

Example 5.3 *Inheritance with exceptions; Inheritable Relations.*

Touretzky [77] identifies two difficulties with multiple inheritance hierarchies when exceptions are allowed. One arises from the presence of redundant information, and the other from the possibility of ambiguity. His examples will be used to demonstrate how the mechanism of belief formation and **Lp** can deal with these problems.

For example, we may have the following information: "Elephants are gray", "Royal Elephants are elephants", "Royal Elephants are not gray", "*Clyde* is a Royal Elephant", and then we add the redundant statement "*Clyde* is an Elephant". Since *Clyde* is a special type of elephant, a type which are not usually gray, we do not want the information that he is also an elephant to trigger an inference that he is gray. That is, we wish *Clyde* to inherit properties from his most specific class. The preference criterion of the belief formation mechanism allows just that.

This knowledge can be encoded in **Lp** with the following set of sentences:

$$
\begin{aligned}
\{ \quad & [Gray(x) | Elephant(x)]_x > \mathbf{c}, \\
& \forall x Royal\_Elephant(x) \rightarrow Elephant(x), \\
& [Gray(x) | Royal\_Elephant(x)]_x < 1 - \mathbf{c}, \\
& Royal\_Elephant(Clyde), Elephant(Clyde) \quad \},
\end{aligned}
$$

where $\mathbf{c}$ is some field constant close to one.[3] Given this knowledge, we have that

---

[3] There is an explicit semantic difference between defeasible properties like "Elephants

$$\mathcal{B}(Gray(Clyde)|Royal\_Elephant(Clyde)) < 1 - \mathbf{c},$$

while

$$\mathcal{B}(Gray(Clyde)|Elephant(Clyde)) > \mathbf{c}.$$

However, the knowledge base can prove that

$$\forall x\ Royal\_Elephant(x) \rightarrow Elephant(x).$$

Hence, $\mathcal{B}(Gray(Clyde)|Royal\_Elephant(Clyde))$ is a preferred degree of belief. It is less than $1 - \mathbf{c}$; hence, it is probable that *Clyde* is not gray.

If it was not known that *Clyde* is a Royal elephant, just that he is an elephant, then belief formation would assign a degree greater than $\mathbf{c}$ to the belief that *Clyde* is gray, given that he is an elephant. If the new information that *Clyde* is a Royal elephant is now added to the knowledge base, then this old degree of belief would be superceded. That is, a preferred degree of belief, based on the knowledge that *Clyde* is a Royal elephant, would now be obtainable from the knowledge base. This is an example of non-monotonic behaviour.

Ambiguity arises in the Nixon example. Here we are given the information "Quakers are Pacifists", "Republicans are not Pacifists", "*Nixon* is

---

are gray" and necessary properties like "Royal Elephants are Elephants", a difference which is expressible in **Lp**, but not in Touretzky's inheritance net formalism.

a Quaker", and "*Nixon* is a Republican." This knowledge could be represented in **Lp** with the following sentences:

$$\{ \quad [Pacifist(x)|Quaker(x)]_x > \mathbf{c}, \quad [Pacifist(x)|Republican(x)]_x < 1 - \mathbf{c}, \quad Quaker(Nixon), Republican(Nixon) \quad \}.$$

Two degrees of belief can be generated,

$$\mathcal{B}(Pacifist(Nixon)|Quaker(Nixon))$$

which is $> \mathbf{c}$, and

$$\mathcal{B}(Pacifist(Nixon)|Republican(Nixon))$$

which is $< 1 - \mathbf{c}$. In this case no preference is deducible between these different degrees of belief, i.e., neither

$$\forall x\ Quaker(x) \rightarrow Republican(x) \text{ nor } \forall x\ Republican(x) \rightarrow Quaker(x).$$

Hence, ambiguity exists, since there are conflicting degrees of belief about *Pacifist*(*Nixon*) and no choice between them.

In this case, since the knowledge base is so small, it is easy to see that true ambiguity exists. However, when the knowledge base is large, it will in general not be known if ambiguity really exists, or if we simply have not yet been able deduce a preference which would resolve the ambiguity. This is a result of the undecidability of **Lp**; i.e., there exists no decision procedure which will detect the non-deducibility of a formula. In this case, the most reasonable action is to be conservative by assuming that the ambiguity is factual.

If later we learn something about the class of people who are both Republicans and Quakers, e.g., $[Pacifist(x)|Quaker(x) \wedge Republican(x)]_x < 1 - \mathbf{c}$, then a new preferred degree of belief would be

$$\mathcal{B}(Pacifist(Nixon)|Quaker(Nixon) \wedge Republican(Nixon))$$

whose value is less than $1 - \mathbf{c}$. This kind of an update is not possible in Touretzky's system, since complex classes formed from conjunctions or other logical connectives are not expressible in his system.

Touretzky also introduces inheritable relations, as an extension of his work on inheritable properties. For example, given the information "Elephants love Zookeepers", "*Clyde* is an Elephant", "*Fred* is a Zookeeper", one would want to conclude that *Clyde* probably loves *Fred*. These kinds of inferences are also possible through belief formation.

This information could be represented in **Lp** as

$$\begin{aligned} \{ \quad & [Loves(x,y)|Elephant(x) \wedge Zookeeper(y)]_{\langle x,y \rangle} > \mathbf{c}, \\ & Elephant(Clyde), Zookeeper(Fred) \quad \}. \end{aligned}$$

From which the degree of belief

$$\mathcal{B}(Loves(Clyde, Fred)|Elephant(Clyde) \wedge Zookeeper(Fred))$$

would be $> \mathbf{c}$.

Inheritable relations can also have exceptions, and these exceptions are handled by belief formation. For example, we may have specific information about *Fred* or *Clyde* which make them exceptions to the generalization "Elephants love Zookeepers"—maybe *Fred* is bad tempered and not loved by many elephants. In this case, the specific information about *Fred* would take precedence.

For example, if the sentence $[Loves(x, Fred)|Elephant(x)]_x < 1-\mathbf{c}$ was added to the knowledge base, then a new preferred degree of belief can be generated in the assertion $Loves(Clyde, Fred)$ which is $< 1-\mathbf{c}$, making it unlikely that *Clyde* loves *Fred*.

When $[Loves(x, Fred)|Elephant(x)]_x < 1-\mathbf{c}$ is added to the knowledge base, the new sentence

$$[Loves(x, Fred)|Elephant(x) \wedge Zookeeper(Fred)]_x < 1-\mathbf{c}$$

can be deduced. (x is not free in $Zookeeper(Fred)$, so this term has the same value as the previous term, by lemma 5.1.1). Hence,

$$\begin{aligned}&\mathcal{B}(Loves(Clyde, Fred)|\\&\quad Elephant(Clyde) \wedge Zookeeper(Fred) \wedge\\&\quad [Loves(x, Fred)|Elephant(x) \wedge Zookeeper(Fred)]_x < 1-\mathbf{c}).\end{aligned}$$

is well founded. Furthermore it is a preferred degree of belief, as it is founded on more knowledge, and its value is $< 1-\mathbf{c}$.

To see that it is $< 1 - \mathbf{c}$, consider the probability term arising from its evaluation:

$$\begin{aligned}&[\mathit{Loves}(z,y)|\\ &\quad[\mathit{Loves}(x,y)|\mathit{Elephant}(x) \wedge \mathit{Zookeeper}(y)]_x < 1 - \mathbf{c} \wedge\\ &\quad\mathit{Elephant}(z) \wedge \mathit{Zookeeper}(y)]_{\langle z,y\rangle}.\end{aligned}$$

This probability term is a variant (definition 4.1.6) of the probability term

$$\begin{aligned}&[\mathit{Loves}(z,y)|\\ &\quad[\mathit{Loves}(z,y)|\mathit{Elephant}(z) \wedge \mathit{Zookeeper}(y)]_z < 1 - \mathbf{c} \wedge\\ &\quad\mathit{Elephant}(z) \wedge \mathit{Zookeeper}(y)]_{\langle z,y\rangle}.\end{aligned}$$

in which the variable x has been replaced by the variable z. By theorem 4.1.7, variants define the same underlying set of objects; thus, these terms have the same value. An application of axiom P9 (along with the equality axioms to take care of the fact that the predicate mentioned is $<$ not $\leq$, and the fact that $0 \leq$ all probability terms) shows that this probability term is $< 1 - \mathbf{c}$. Hence, the original term is $< 1 - \mathbf{c}$, and the degree of belief is $< 1 - \mathbf{c}$, as claimed.

This example shows that the same preference criterion can be used for inheritable relations as well.

There has been a lot of recent work on inheritance systems which allow exceptions, most of it based on graphical structures called inheritance nets.

These structures are an example of the specialized structures which can be used to speed up reasoning, mentioned above, in section 5.1.3. The next chapter shows how such a structure can organize the kind of knowledge used in the first part of this example, i.e., universal set inclusions, like "Royal elephants are elephants," and statistical generalizations, like "Most elephants are gray." This structure permits the rapid deduction of the values of the degrees of belief in assertions which concern property inheritance.

Chapter 6

An Inheritance Reasoner

The problem of reasoning about property inheritance in multiple inheritance hierarchies when exceptions are allowed, has received a lot of recent attention in AI (e.g., Touretzky [77], Horty et al. [30], Touretzky et al. [78], Pearl [57], Geffner and Pearl [23], Etherington [19], Sandewall [69]). The previous chapter presented an example which demonstrated that the main problems arising from the existence of exceptions, i.e., ambiguity and redundancy, can be dealt with by the formalism developed in this thesis. The example used unstructured sets of **Lp** sentences to represent the knowledge and belief formation, along with **Lp**-deduction, to infer the plausible property inheritances.

In contrast, most other works on inheritance systems (including most of the work cited above) have represented the knowledge in a highly structured graph formalism called an inheritance net, and has performed inference by finding paths in this net. This chapter demonstrates that such a structured representation is also possible within the formalism of **Lp** and belief formation.

Structured representations can be viewed as being mechanisms for organizing knowledge. Efficient structured representations are available for certain sets of knowledge, but a large knowledge base will also require a more general scheme for representing knowledge. Since a general scheme must be provided, it is profitable to view all of the knowledge in the knowledge base as being expressed in the same general scheme. This eliminates the need for providing separate semantics for each different part of the knowledge base. In this view, the structured representations are not different formalisms for representing knowledge; they are instead structures for organizing knowledge represented in some underlying general formalism. The chief advantage of these organizational structures is that they allow inferences to be generated through simple procedures, e.g., graph traversal, instead of through the application of an abstract, and probably computationally complex, inference procedure, e.g., a proof theory. Thus, another way of viewing these organizational structures is as a compilation of the knowledge; i.e., the form of the knowledge is changed to increase efficiency, but its meaning remains intact.

In this chapter inheritance graphs are viewed as being organizational structures for a special subset of knowledge, the subset of general knowledge which is commonly used for reasoning about property inheritance. It will be demonstrated how this knowledge, expressed in **Lp**, can be structured into an inheritance net. This net has the property that traversing certain types of paths in it is equivalent to performing certain deductions in **Lp**. The net permits the rapid inference of plausible conjectures about an individual's properties. These plausible conjectures take the form of closed formulas which assert that a particular individual has a particular property and which can be given a high degree of belief by belief formation.

6.1 The Nature of Inheritance Systems

The Knowledge

Inheritance systems are concerned with classes which are sets of individuals with a particular property, and with individuals who are members of some set of classes, e.g., the class of elephants which is the set of individuals who are elephants (i.e., they possess the property "elephantness"), the class of gray which is the set of gray individuals, or *Clyde* a particular individual who is an elephant (i.e., a member of the class of elephants). The members of certain classes are known to be members of other classes, either uniformly or with allowance for exceptions, e.g., the members of the class of elephants are also members of the class of mammals, without exception, or the members of the class of birds are usually members of the class of flying objects, but there are exceptions.

Hence, there are three types of knowledge present in an inheritance system:

1. some set of known individuals and known classes;
2. some set of assertions which state that particular individuals are members of particular classes;
3. some set of assertions which state that the members of a certain class have some property, i.e., they are also members of some other class, either uniformly or with exceptions.

Most systems extend this set of knowledge types by also allowing items 2 and 3 to be negative claims. That is, they allow the claims that a particular individual is not a member of a particular class, and that members of a certain class do not have some property.

In inheritance systems which represent this knowledge as an inheritance net, item 1 is represented as a set of nodes in the net, item 2 is represented as links between nodes representing individuals to nodes representing classes, and item 3 is represented as links between nodes which represent classes.

In **Lp** this knowledge can be represented as: item 1—object constant symbols for the individuals and object predicate symbols for the classes;

item 2—atomic formulas formed by applying the appropriate predicate to the appropriate constant; and item 3—in the case of uniform assertions, as universal sentences formed with the appropriate pair of predicates, and in the case of exception allowing assertions, as probability terms formed from the predicates. For example, we may have the knowledge:

1. *Clyde* and *Tweety* are individuals; *Elephant*, *Animal*, *Bird*, *Gray*, and *Fly* are classes;
2. *Clyde* is an *Elephant*, and *Tweety* is a *Bird*;
3. *Birds* and *Elephants* are *Animals* without exception; *Birds* can *Fly*, and *Elephants* are *Gray*, with exceptions.

This knowledge can be encoded in **Lp** as the following:

1. *Clyde* and *Tweety* are encoded as object constant symbols, and each of the classes are encoded as object predicate symbols;
2. $Elephant(Clyde)$ and $Bird(Tweety)$ are atomic formulas;
3. $\forall x\ Bird(x) \rightarrow Animal(x)$, $\forall x\ Elephant(x) \rightarrow Animal(x)$, $[Fly(x)|Bird(x)]_x > \mathbf{b}$, and $[Gray(x)|Elephant]_x > \mathbf{b}$, where '**b**' is some field constant close to one.

Chapter 5, example 5.3, gives a few more examples of this encoding. It will be shown in this chapter how the set of **Lp** formulas produced by this encoding can be mapped onto an inheritance net which has two types of links, in such a manner that there is a one-one correspondence between the nodes and links in the net and the **Lp** formulas.

The Inferences

The knowledge in the inheritance system is used to infer the properties of various individuals. Each individual in the system is known to be a member of certain classes, and using the information present about the relationships between classes, other properties (i.e., class memberships) can be inferred. For example, since in the above example all elephants are animals, it can be inferred that *Clyde* the elephant is also an animal, or since most elephants are gray, it can be inferred that *Clyde* is probably gray.

It has previously been demonstrated, in chapter 5, how **Lp** deduction and belief formation can be used to produce these inferences. In graph based approaches, these inferences are produced by tracing paths which emanate from the node representing the individual and terminate at nodes representing other properties. For example, in a graph representing the above set of knowledge (see, figure 1), the node *Clyde* would be connected to an "elephant" node which would in turn be connected to an "animal" node. By tracing the path from *Clyde* to the "animal" node, a graph based system could produce the inference that *Clyde* is an *Animal*.

6.2 Heterogeneous vs. Homogeneous Inheritance Systems

Using the terminology of Touretzky et al. [78], inheritance system can be divided into homogeneous and heterogeneous systems. Most of the work on inheritance systems has concentrated on homogeneous system (including, Touretzky [77], Horty et al. [30], Pearl [57], Sandewall [69]). Homogeneous systems do not differentiate relationships between classes which hold uniformly from relationships which allow exceptions. So, for example, a homogeneous system would not differentiate between the assertions "Elephants are mammals" and "Most birds fly", even though the former allows no exceptions while the latter does. In order to model defeasible (exception allowing) properties of classes, homogeneous systems represent all properties of classes as defeasible properties, dispensing with necessary properties. Heterogeneous systems, on the other hand, do make an explicit differentiation between necessary and defeasible properties of classes. Heterogeneous systems have been proposed before, e.g., in Etherington [18] and more recently in Geffner and Pearl [23]. The system presented here is heterogeneous.

Homogeneous inheritance systems have been criticized at length by Brachman in [5]. The main focus of his very sound criticisms is that since there is no differentiation between defeasible and necessary properties, all conclusions drawn by homogeneous systems must be defeasible. That is, it is always possible that the system may be lying. This makes it impossible for such systems to represent compositional classes accurately, e.g., classes like three legged elephants. One could never conclude with certainty that three legged elephants possess three legs, even through this follows from the definition.

Homogeneous systems also exhibit another difficulty not mentioned by Brachman. From a formal standpoint there is no a priori limit on the depth of an inheritance net, and thus, no a priori limit on the length of paths down which properties can be inherited. There is no problem with inheritance down an arbitrarily lengthy path of strict IS-A links. For example, given the strict IS-A path

$$Tweety \Rightarrow Bird \Rightarrow Animal \Rightarrow Physical\ Object \Rightarrow Occupies\ Space,$$

the inference that Tweety occupies space is perfectly valid and intuitive. However, property inheritance down a path of defeasible links can quickly lead to counter-intuitive conclusions. For example, the path of defeasible links

$$Helicopter \rightarrow Flying\ Object \rightarrow Has\ Wings$$

leads to the counter-intuitive conclusion that helicopters have wings after only two perfectly reasonable defeasible inferences. It might be argued that in any inheritance net the node ***Helicopter*** should have an exception link to the node *Has Wings*. However, it is easy to see that in general avoiding all such counter-intuitive conclusions would necessitate examining all allowable paths in the inheritance net and then adding the required exception links. I would argue that this vitiates the very purpose of inheritance systems. Inheritance systems, like any reasoning system, are intended to generate plausible conclusions which go beyond the explicit knowledge actually contained in them. To require the addition of such intuition preserving exception links is, in a sense, requiring that the conclusions be known prior to any reasoning being performed.

Homogeneous systems, through their inability to differentiate between necessary and defeasible links, cannot know when a lengthy chain of inheritance leads to a valid conclusion and when it leads to a counter-intuitive conclusion. If lengthy chains are prohibited in an attempt to avoid counter-intuitive conclusions, many valid chains will also be excluded. If lengthy chains are allowed, there will be some counter-intuitive conclusions generated.

The heterogeneous system of Geffner and Pearl [23] also suffers from this difficulty. They interpret defeasible links as holding with probability almost one. Hence, their system sanctions property inheritance down arbitrarily lengthy chains of defeasible links.

6.3 Representing Statements of Typicality With Probabilities

It can be seen from the example in section 6.1 that the proposed encoding of defeasible properties in **Lp** is as high valued conditional probability terms. For example, the defeasible property "birds fly" will be encoded by asserting that the probability term $[Fly(x)|Bird(x)]_x$ has a value close to one. Semantically, this means that we are encoding the defeasible property in a statistical manner; i.e., as "A large percentage of birds can fly."

It has, however, been claimed that probabilities are inappropriate to encode notions of typicality, and for many researchers inheritance systems are intended to reason with notions of typicality. The example of "Dogs give live birth", used by Carlson [7], points out difficulties with a probabilistic interpretation of statements of typicality (see also a recent article by Nutter [54]). Giving live birth is a typical property of dogs (or any mammal) which is not, however, a property possessed by a majority of dogs. This example demonstrates the need for a careful distinction between different notions of typicality.

Brachman [5, page 98] has pointed out that there is a difference between prototypical properties, which, in a sense, are characteristic of a kind, and descriptions which specify the properties which usually apply to instances of a kind. It is certainly true that prototypical properties may not be probable properties and as such not representable in the semantic model constructed here.[1] However, Brachman argues that inheritance nets are more concerned with the expression of properties which usually apply to instances of a kind, rather than prototypical properties. Indeed, given an instance of a kind the inheritance net is used to reason about the properties which that instance may possess. Surely, the fact that a large percentage of instances of that kind possess a property is a good reason to conjecture that a particular instance possesses that property.

In fact, it is not clear that inheritance systems are appropriate for reasoning about prototypical properties. In the typical use of inheritance systems we are given an individual who is known to possess some set of properties, e.g., Tweety who is a bird and a canary. The task of the inheritance system is to generate a new set of plausible properties which that individual may also possess, based on the properties which he is known to have; e.g., since Tweety is a bird it is plausible that he can fly. If prototypical properties like "birds lay eggs" were to be encoded in the inheritance net, one would be led to rather counter-intuitive conclusions like "Tweety lays eggs," once it was known that Tweety is a bird. The problem here is that it is not always reasonable to conclude that an arbitrary individual is prototypical. There may be situations where the assumption of prototypicality is reasonable, but it could be claimed that such an assumption is hardly ever reasonable in the advertised use of inheritance systems.

---

[1] This does not necessarily mean that probabilities are not useful for expressing notions of prototypicality. For example, although the majority of dogs do not give live birth the probability of a dog giving live birth is much greater than the probability of, say, a bird giving live birth. So, perhaps prototypical properties can be expressed by ratios of probabilities. If they were so expressed, the probability ratios could be used to reason with these properties. For example, if we know that *Fido* is either a dog or a canary, and then we are told that she had given birth to a live litter, a simple Bayesian model would allow one to make the reasonable conclusion that *Fido* is probably a dog.

Pearl [57] has also proposed a probabilistic interpretation for defeasible properties. In his system a defeasible property is encoded as a property which has probability $1 - \epsilon$, where ϵ can be made arbitrarily close to zero. However, his is a homogeneous system; hence, all assertions are assumed to hold with high probability, even certain assertions like "Tweety is a bird" or "Birds are animals." Even without considering the problems which arise from homogeneity, the assertion that defeasible properties have an arbitrarily high probability is counter-intuitive. It may be true that a large percentage of birds can fly, but this percentage is surely not *arbitrarily* close to one hundred. The system of Geffner and Pearl [23] is also subject to this criticism; they interpret defeasible properties as holding with probability approximately one. The interpretation offered here, i.e., that defeasible properties hold with probability greater than some constant which is close to one seems to be more natural. It has not been specified how close to one this constant is, but it is not arbitrarily close.

One possible problem with this interpretation of defeasibility is that properties which are probable cannot serve as the basis for further conjecture. In the graph representation, this means that only one defeasible link can be traversed in any path. It is easy to see that even if more than 100**b**% of all P's are Q's and more than 100**b**% of all Q's are S's there is no reason to suppose that more than 100**b**% of all P's are S's. In fact, there is no reason to suppose that even more that 50% of all Q's are S's. If the set of Q's is much larger than the set of P's we could have 99% of all P's being Q's and 99% of all Q's being S's and still have *no* P's being S's (try 100 P's and 10,000 Q's). Property inheritance down more than one defeasible link can never be *uniformly* valid within these semantics. In fact, this is the reason why chains of defeasible links sometimes lead to counter-intuitive conclusions.

There are certain situations where it seems intuitive to allow property inheritance down more that one defeasible link, and this can be done with an additional independence assumption. For example, if the conditional probability of S given Q is independent of P, i.e., $[S(x)|Q(x) \wedge P(x)]_x = [S(x)|Q(x)]_x$, then it is deducible that $[S(x)|Q(x) \wedge P(x)]_x \geq [S(x)|Q(x)]_x \times [Q(x)|P(x)]_x$; $[S(x)|P(x)]_x \geq [S(x) \wedge Q(x)|P(x)]_x = [S(x) \wedge Q(x)|P(x)]_x \times \frac{[Q(x)|P(x)]_x}{[Q(x)|P(x)]_x} = [S(x)|Q(x) \wedge P(x)]_x \times [Q(x)|P(x)]_x = [S(x)|Q(x)]_x \times [Q(x)|P(x)]_x$. Thus, with this independence assumption it can be deduced that $> (100)\mathbf{b}^2\%$ of all P's are S's. This has the natural result that each time a defeasible property is used the final conclusion becomes less certain. If the probability of each defeasible property is very high, i.e., **b** is close to one, then the final conclusion will have a reasonably high probability. The real problem, however, is deciding under what situations this assumption of independence is valid. Clearly such an assumption is not uniformly valid, as was indicated by the "helicopter has wings" example.

It will be seen later that limiting the use of defeasible properties does not exclude lengthy chains of inheritance. Property inheritance can occur down arbitrary lengthy chains, but all of the links in the chain will be strict IS-A links except, possibly, for one defeasible link. However, this still means that some of the examples which have appeared in the literature cannot be handled by this reasoner. As a result, it could be claimed that this reasoner is incomplete; however, the motivation behind this work is to develop an inheritance reasoner which is sound. That is, one that will not produce invalid inferences irrespective of the meaning of the nodes in the net.

6.4 The Inheritance Graph

The formal details of the inheritance reasoner are presented in this section. First, the set of **Lp** sentences which can be used as knowledge is specified more precisely, as follows:

1. We have a set of object constant symbols, e.g., *Tweety, Fido, Nixon*, and a set of unary[2] object predicate symbols, e.g., *Bird, Fly, Elephant*,

---

[2] Relations are not handled by this reasoner.

Animal, and finally, a field constant symbol, b, which is in the open interval (0.5, 1).

2. There is a set of atomic formulas and negations of atomic formulas constructed from the above set of predicate and constant symbols, e.g., $Bird(Tweety)$, $\neg Fly(Tweety)$.

3. Finally, there is a set of universal implications formed from the above predicate symbols, e.g., $\forall x\ Bird(x) \rightarrow Animal(x)$, where the consequent predicate may be negated, e.g., $\forall x\ Elephant(x) \rightarrow \neg Bird(x)$, and a set of probability assertions containing probability terms formed from the above predicate symbols, e.g., $[Fly(x)|Bird(x)]_x > \mathbf{b}$, these assertions can also be negated, e.g., $[Fly(x)|Elephant(x)]_x < 1 - \mathbf{b}$.

An inheritance graph can be constructed which encodes this knowledge. The graph has four types of four link symbols, $\Rightarrow$ (IS-A), and $\not\Rightarrow$ (IS-NOT-A), which are strict links, $\rightarrow$ (Probably-IS-A), and $\not\rightarrow$ (Probably-IS-NOT-A) which are defeasible links. Each constant and predicate symbol generates a node in the graph. The atomic formulas generate strict links running from the node representing the individual to the appropriate property node. The universal implications generate strict links between the two property nodes, while the probability assertions generate defeasible links between the property nodes, IS-NOT-A links are used when the assertions are negated. For example, the knowledge given in section 6.1 generates the

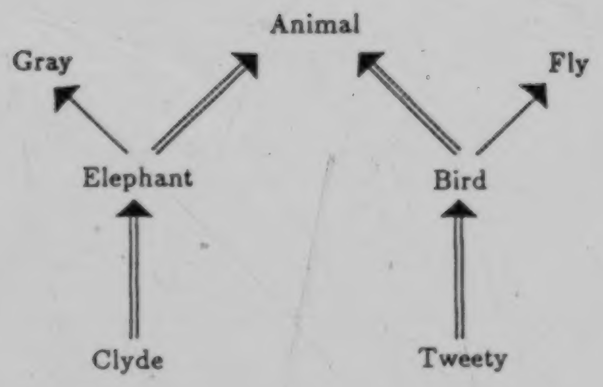

Figure 1: The Graph Encoding

graph given in figure 1.

Inference about the properties of a particular individual is accomplished by following certain types of paths in the graph which emanate from the node representing that individual. All of the property nodes which are reachable from that node through legal non-preempted paths can be concluded to be possible properties of the individual.

The next section defines what constitutes legal paths, as well as what it means for a path to be preempted.

6.5 Inferences in the Inheritance Net

Reasoning in the inheritance net is performed by finding paths, and exceptions are handled by a concept of path preemption. These are notions derived from Touretzky's original work on inheritance hierarchies [77]. First, we define the legal paths and the conclusions they support, then path preemption, and finally, once the interaction between paths is taken into consideration, the conclusions supported by the inheritance graph as a whole.

There are two types of paths, necessary paths and probable paths. The following definitions are based on finding paths emanating from the node representing a particular individual c. Furthermore, P is used to represent a property node, a superscripted plus sign (i.e., $^+$) indicates one or more iterations of the link to which it is applied, and a superscripted star (i.e., *) indicates zero or more iterations of the link to which it is applied.

Definition 6.5.1 (Paths) *Positive and Negative, Necessary and Probable paths are defined as follows:*

Positive Necessary 1 A path $c \Rightarrow^+ P$ supports the conclusion *c is certainly a P.*

Negative Necessary 1 A path $c \Rightarrow^* \bullet \not\Rightarrow \bullet \Leftarrow^* P$ supports the conclusion *c is certainly not a P.*

Negative Necessary 2 A path $c \Rightarrow^{+} \bullet \nLeftarrow \bullet \Leftarrow^{*} P$ supports the conclusion c *is certainly not a* P.

Positive Probable 1 A path $c \Rightarrow^{+} \lambda \rightarrow \bullet \Rightarrow^{*} P$ supports the conclusion *probably* c *is a* P *founded on* λ. That is, this conclusion is based on c being an random member of the class λ.

Negative Probable 1 A path $c \Rightarrow^{+} \lambda \not\rightarrow \bullet \Leftarrow^{*} P$ supports the conclusion *probably* c *is not a* P *founded on* λ.

Negative Probable 2 A path $c \Rightarrow^{+} \lambda \rightarrow \alpha \cdots R$, where $\cdots$ represents a negative necessary path (either type 1 or 2) from α to R, supports the conclusion *probably* c *is not a* P *founded on* λ.

As long as the original set of **Lp** sentences is consistent, they will have a model, by theorem 4.2.15. This means that there will be some set of objects for which the assertions are true. We impose the restriction that the original set of **Lp** sentences be consistent, and say that the net is consistent if and only if the original set of **Lp** sentences are. For the net to be consistent it cannot contain both a necessary positive path and a necessary negative path from any individual to the same property node, nor can it contain both a probable positive and a probable negative path from any individual c to the same property node *founded on the same class*. Consistency does not require, however, that the graph be acyclic. In fact two properties may often be cross correlated (i.e., existence for either may provide evidence for the other). The restriction to acyclic graphs has been cited as a deficiency of previous inheritance reasoners (Geffner and Pearl [23]). Figure 4 presents an example of a graph representing a reasonable set of knowledge which is not acyclic.

Definition 6.5.2 (Path Preemption) *Probable paths are preempted if any of the following conditions hold:*

1. A probable path from c to P is *preempted* if there is a necessary path from c to P. In this case the polarity of the two paths is irrelevant.
2. A probable path from c to P founded on λ_1 is *preempted* if there is a probable path of the opposite polarity from c to P founded on λ_2 such that there exists a path $\lambda_2 \Rightarrow^+ \lambda_1$ in the graph.
3. A probable path from c to P is *preempted* if any of its subpaths are preempted.

Definition 6.5.3 (Conclusions Supported by the Graph) *Given an individual c, the graph supports the following conclusions:*

1. c *is certainly a* P if a positive necessary path exists from c to P.
2. c *is certainly not a* P if a negative necessary path exists from c to P.

3. *c is evidently a P* if there exists a non-preempted positive probable path from c to P, and there does not exist any non-preempted negative probable paths from c to P.

4. *c is evidently not a P* if there exists a non-preempted negative probable path from c to P, and there does not exist any non-preempted positive probable paths to from c to P.

5. The graph is *ambiguous* about c being a P if there exists both a non-preempted positive probable and a non-preempted negative probable path from c to P.

6.6 The Relation to Belief Formation

All of the certain conclusions supported by the graph are deductive consequences of the original set of knowledge. That is, given the original set of **Lp** sentences the certain conclusions can be deduced using **Lp** deduction. Each link in one of the certain paths represents a universal implication, except for the first link (from the individual node to a property node) which represents an atomic formula. For the positive necessary paths, it is easy to see that the final consequence of the chain of universal implications can be deduced given the instantiation of the initial antecedent. For example, given the chain $\forall x\, P(x) \rightarrow Q(x)$, $\forall x\, Q(x) \rightarrow R(x)$, and an instantiation of the initial antecedent, $P(c)$, $R(c)$ can be deduced. The negative necessary paths can also be seen to yield deductive consequences once one notes that $\forall x \neg Q(x) \rightarrow \neg P(x)$ is a tautological consequence of $\forall x\ P(x) \rightarrow Q(x)$.

The probable paths yield conclusions supported by the mechanism of belief formation. Since the path to the founding class λ is a positive necessary path, λ is a deductive property of the individual c, i.e., $\lambda(c)$ can be deduced from the original set of **Lp** sentences. For the positive probable paths the probable link from λ to the next node in the path, call it α, encodes the probability assertion $[\alpha(x)|\lambda(x)]_x > \mathbf{b}$; hence, belief formation gives a value $> \mathbf{b}$ to the degree of belief $\mathcal{B}(\alpha(c)|\lambda(c))$. That is, based on the knowledge $\lambda(c)$, $\alpha(c)$ can be held to a high degree, $> \mathbf{b}$. Since α is connected to the final node in the path via a positive necessary path, $\forall x\, \alpha(x) \rightarrow P(x)$ can be deduced. Lemma 4.3.3 shows that $[P(x)|\lambda(x)]_x$ is larger than $[\alpha(x)|\lambda(x)]_x$; thus the degree of belief $\mathcal{B}(P(c)|\lambda(c))$ is greater than $\mathbf{b}$. Similar reasoning shows that the negative probable paths also generate conclusions supported by belief formation.

The conclusions supported by these paths are only based on the knowledge $\lambda(c)$. However, $\lambda(c)$ is, in general, not the only knowledge available about the individual c. For example, in the Nixon diamond, figure 3, both $Republican(Nixon)$ and $Quaker(Nixon)$ are known properties of the individual $Nixon$. As was shown in chapter 5, different knowledge can generate different degrees of belief in an assertion. In the Nixon diamond, there are two probable paths from ***Nixon*** to ***Pacifist***, a positive one founded on ***Quaker***, and a negative one founded on ***Republican***. These two paths correspond to the two degrees of belief $\mathcal{B}(Pacifist(Nixon)|Quaker(Nixon))$, which is greater than **b**, and $\mathcal{B}(Pacifist(Nixon)|Republican(Nixon))$, which is than $1 - \mathbf{b}$. In this net neither path preempts the other; thus, the net is ambiguous about Nixon being a pacifist. This corresponds to the situation where there is no preference between the different degrees of belief. A preference criterion was, however, presented in chapter 5. This criterion is reflected as path preemption in the graph based reasoner.

A probable path preempts a probable path of the opposite polarity if there is a positive necessary path from its founding class, λ_1 to the founding class of the preempted path, λ_2. The presence of a positive necessary path indicates that $\forall x\ \lambda_1(x) \rightarrow \lambda_2(x)$. Since the probable paths correspond to the two degrees of belief $\mathcal{B}(P(c)|\lambda_1(c))$, and $\mathcal{B}(P(c)|\lambda_2(c))$, it can be seen that the preference criterion sanctions the preemption of the path based on the knowledge $\lambda_2(c)$.

A probable path is also preempted if any of its subpaths are preempted. For example, say we have a positive probable path $c \Rightarrow \lambda_1 \rightarrow R \Rightarrow P$ and a negative probable path $c \Rightarrow \lambda_2 \not\rightarrow R$, where λ_2 is a subset of λ_1. In this case the positive path is preempted because the subpath from c to R is preempted. Since more than 100**b**% of all λ_1's are R's it is necessarily the case that more than 100**b**% of them are P's. On the other hand, less than

100(1 – **b**)% of all λ_2's are R's, and this gives no reason to conjecture any thing about the proportion of them that are members of the superset P. As noted above, degrees of belief based on the knowledge that c is a λ_2 are preferred to degrees of belief based on the knowledge that c is a λ_1; so, we are left with no conclusions about c possessing the property P.

The existence of non-preempted probable paths of only one polarity from an individual c to a property P indicates that all of the evidence in the net about c's P-ness is of that polarity. It is in this sense that the evidential conclusions are supported by the net.

6.6.1 On the Independence Assumptions Underlying Inheritance Reasoners

The probabilistic semantics of **Lp** make it clear that there are some implicit assumptions being used in inheritance reasoners.

One implicit assumption was previously identified by Pearl [57], who calls it the Principle of Positive Conjunction. In actuality it is a principle of both positive and negative conjunction. The graph will support the conclusion that c is evidently a P if more that one non-preempted positive path exists from c to P. In probability theory it is not necessarily the case that two items of positive evidence remain positive. That is, if $[P(x)|R(x)]_x > \mathbf{b}$ and $[P(x)|Q(x)]_x > \mathbf{b}$ it is not necessarily the case that

$[P(x)|R(x) \wedge Q(x)]_x > \mathbf{b}$. When there is more that one positive (negative) probable path from c to P each path may be based on a different piece of knowledge. In continuing to conclude that c is probably (not) a P, it is being implicitly assumed that such cancellation does not occur. This assumption also exists in the non-probabilistic systems of Touretzky and Horty et al. In fact, because the net is incapable of representing conjunctive classes like $R(x) \wedge Q(x)$, graph based inheritance reasoners cannot help but make this assumption; they have no means of representing any cancellations that might exist. The general mechanism of belief formation, which is performed on unstructured sentences of **Lp**, is capable of representing and taking into consideration cancellation of this form.

Another implicit assumption occurs when no path exists from an item of knowledge to the property we are conjecturing. For example, in multiple inheritance nets Clyde the elephant may also be a circus performer. He will inherit properties from both the class of elephants and the class of circus performers. However, it is assumed that the knowledge that he is a circus performer will not affect properties he may inherit from the class of elephants, if those properties are not connected to the circus performer node. So, for example, the graph may support the conclusion that Clyde is gray if there exists a probable path from Clyde to gray founded on the class elephant, but this conclusion fails to consider the knowledge that Clyde is also a circus performer. This failure occurs because there is no information which indicates that being a circus preformer affects the property of being gray, and it is equivalent to an implicit assumption that the net is complete in the sense that all possible influences are represented in the net.

The existence of these implicit assumptions places restrictions on the generality of graph based inheritance reasoners. There are certainly domains where such reasoners seem to be useful, e.g., animal taxonomies; however, some of the proposed applications are questionable. Identifying applicable domains and more importantly, characterizing the set of applicable domains, remain open problems.

6.7 Behavior of the Reasoner

This section examines the behavior of the reasoner developed in this chapter through the use of some examples. All of the examples have appeared previously in the literature. However, since this reasoner is heterogeneous, the example nets have been altered. In particular, some of the edges in these examples have been changed to strict IS-A links. In some cases it is clear which edges should be written as strict links. For example, Royal elephants are strictly a subset of elephants. In other cases the decisions may be contentious, but nevertheless, the examples still serve to explicate the behavior of this inheritance reasoner.

First, the behavior of the reasoner is examined in the presence of re-

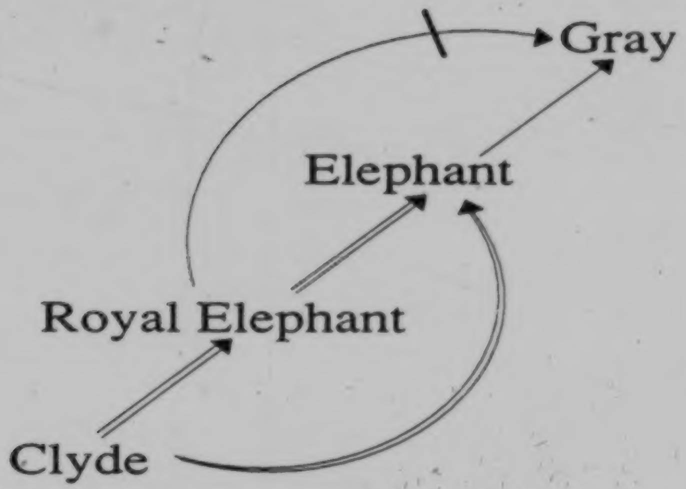

Figure 2: Redundant Information

dundancy and in the presence of ambiguity. These two features of inheritance nets were the motivation for Touretzky's original work on inheritance systems [77]. The gray elephant net is shown in figure 2. In the net a non-preempted negative probable path exists from *Clyde* to the property node *Gray* founded on the property *Royal Elephant*. The opposing positive probable path, founded on the node *Elephant*, is preempted since Royal elephants are a subset of elephants, as indicated by the presence of a $\Rightarrow^{+}$ path from *Royal Elephant* to *Elephant*. Hence, the net sanctions the conclusion that *Clyde* is probably not gray, based of *Clyde* being a royal elephant. An interesting point is that the evidence used to generate the plausible conclusion can be retrieved by examining the path used.

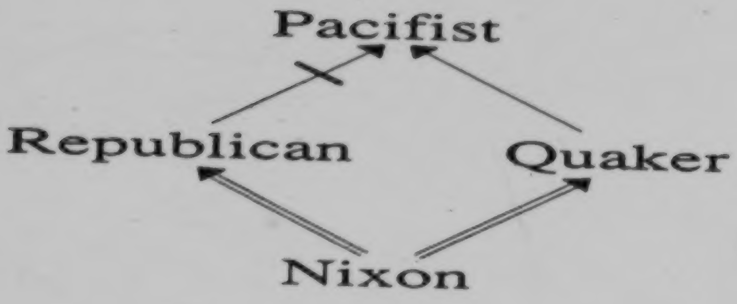

Figure 3: Ambiguous Information

Figure 3 demonstrates the existence of true ambiguity. In this graph, called the Nixon diamond, there is both a positive and a negative probable non-preempted path to the node *Pacifist*, sanctioning the conclusion that the graph is ambiguous. The ambiguity corresponds to the fact that neither of the founding sets for the two paths, *Republican* and *Quaker*, is a subset of the other.

These two examples demonstrate that the main desiderata for inheritance reasoners are satisfied by this reasoner. The next two examples demonstrate some additional features of the reasoner.

The next example, in figure 4, is due to M. Ginsberg. This graph supports the conclusion that *Nixon* is probably politically motivated, but is at the same time ambiguous about *Nixon* being either a dove or a hawk. The knowledge contained in the graph is quite reasonable, and the conclusions which it supports, intuitive. However, it contains a cycle and thus

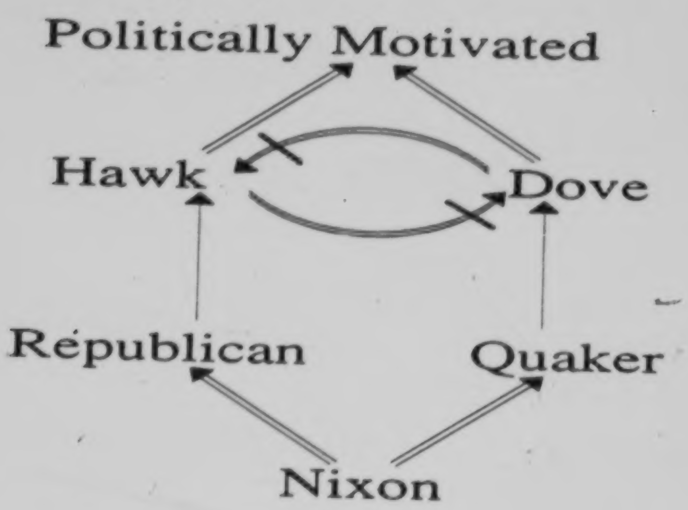

Figure 4: (M. Ginsberg) Is Nixon Politically motivated?

cannot be dealt with by either Touretzky's system [77] or Horty et al.'s skeptical reasoner [30]. The system of Geffner and Pearl [23] can handle this example.

The negative paths generated by backward IS-A links can sometimes generate interesting conclusions, as in figure 5, although, many times the conclusions are uninteresting negative facts.[3] This graph supports the conclusion that *Tweety* is probably not a penguin founded on *Tweety* being a bird (through a negative probable 2 path). Semantically, the conclusion that most ($> 100b\%$) birds are not penguins is entailed by the knowledge

---

[3] As George Bernard Shaw once said "An intelligent man wants to know what you believe, not what you don't believe."

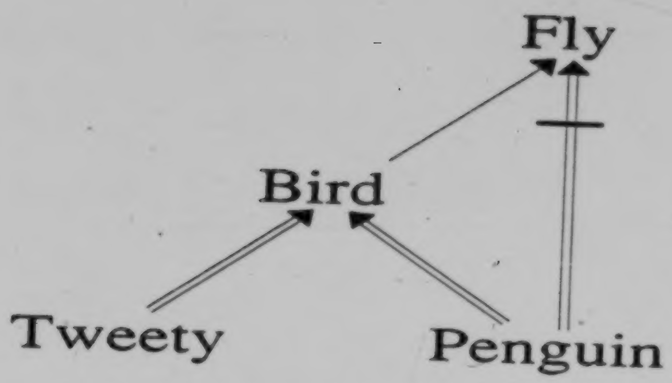

Figure 5: Tweety is probably not a penguin

in this graph. Geffner and Pearl (op. cit.) have argued that the ability to make such inferences, which essentially are a consequence of the properties of probabilities, represents an argument in favor of using probabilities as a semantic foundation for defeasible inference.

6.8 Touretzky et al.'s Design Space

Touretzky et al. [78] present a taxonomy of different design decisions which they claim represent reasonable alternatives for any inheritance reasoner. They present three areas where choices can be made, thus implicitly defining a design space for inheritance reasoners. These choices are discussed in this section, and an attempt is made to identify where in the design space this reasoner lies. Again, the fact that this reasoner is heterogeneous places a caveat on the comparisons. It will be seen, however, that the existence of an underlying set of **Lp** sentences, which make the knowledge contained in the net explicit, sheds some light on the "clashes of intuition" discussed by Touretzky et al.

6.8.1 Skepticism vs. Credulity

The first choice is between skepticism and credulity. A skeptical reasoner tries to be careful in what it concludes, while a credulous reasoner tries to conclude as much as possible. This reasoner can be said to be skeptical, in that all of the conclusions it draws have a semantic justification, given by the justifications presented for belief formation. Furthermore, the conclusions can be 'hedged' by identifying the knowledge on which the probable paths are founded. It is not, however, skeptical in the sense that it refuses to draw conclusions in ambiguous situations, as does the skeptical reasoner presented by Horty et al. [30]. Instead, when there is both positive and negative support for an assertion this reasoner draws the conclusion that the net is ambiguous about that particular assertion. For example in figure 3, Horty et al.'s skeptical reasoner draws no conclusions about *Nixon* being a pacifist, while this reasoner draws the conclusion that the graph is ambiguous about this assertion. It is only when there are no paths from the individual to the property in question that this reasoner draws no conclusions.

In a sense this behavior is similar to the original credulous reasoner of Touretzky [77], except that Touretzky's reasoner generates multiple extensions. In one extension *Nixon* is a pacifist while in another extension *Nixon* is not a pacifist. Ambiguity is detected by the presense of multiple extensions. However, in any particular extension there is no ambiguity, i.e., the assertion is either true or false in that extension. This reasoner does not generate multiple extensions, the conclusion that ambiguity exists about *Nixon* being a pacifist is not expressed by asserting that in one possible model *Nixon* is a pacifist and in another he is not. The reasoner simply asserts that there is evidence for both conclusions.

Horty et al.'s skeptical reasoner turns out to be not completely skeptical; it does not propagate ambiguity. Figure 6a shows a net with cascaded ambiguities taken from Touretzky et al. [78] and altered to be heterogeneous. In the homogeneous version of this net (just make all of the links the same type while preserving polarity) Horty et al.'s reasoner makes no claims about *Nixon* being a pacifist, but it concludes that *Nixon* is not anti-military. This behavior is a result of it being skeptical about pacifism. Since there is no path to *Pacifist*, there is no opposing positive path to *Anti-military*. This reasoner does not propagate ambiguity either, although in certain heterogeneous topologies it behaves as if it did. Figure 6b shows a heterogeneous topology (which would be identical to figure 6a in a homogeneous system) in which the reasoner behaves as if it was propagating

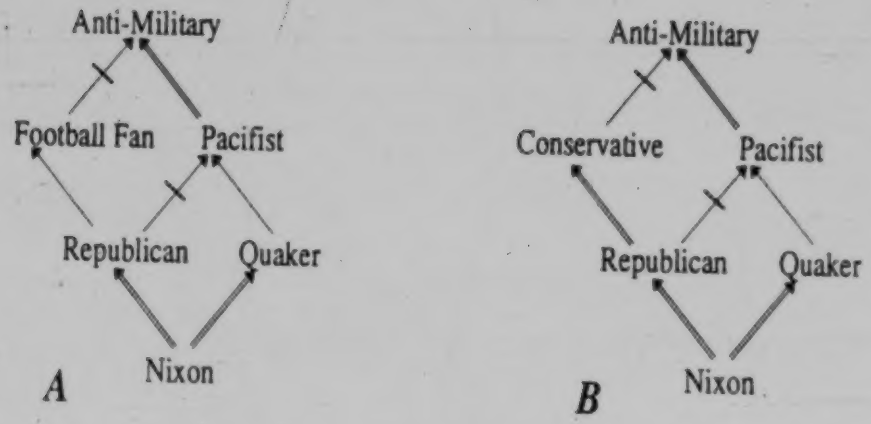

Figure 6: Two examples of cascaded ambiguities

ambiguity. This graph is ambiguous both about *Nixon* being a pacifist and about *Nixon* being anti-military.

In this example, it is illuminating to examine the knowledge contained in the net. It is not accurate to say that ambiguity is being propagated. In fact, one could remove the probable link from ***Republican*** to ***Pacifist*** and the graph would become unambiguous about *Nixon* being a pacifist while remaining ambiguous about *Nixon* being anti-military. If one examines the meaning of the edges in this net it is clear that it contains *no* direct information about the anti-military tendencies of Republicans. It just contains information about the anti-military tendencies of conservatives and Quakers. The positive probable path to anti-military is based on *Nixon* being a

Quaker, while the negative probable path is based on *Nixon* being a conservative. The fact that *Nixon* is a Republican has no influence on him being anti-military. For example, *Nixon* could be a non-Republican conservative and the graph would still support the probable path to anti-military founded on *Nixon* being a conservative.

Figure 6a further illustrates that ambiguity is not being propagated by this reasoner. When reasoning about *Nixon* in the net in this figure the reasoner concludes that *Nixon* is probably a football fan (based on him being a Republican), that the net is ambiguous about *Nixon* being a pacifist, and that *Nixon* is probably anti-military based on him being a Quaker. The behavior with respect to the last conclusion is the opposite of Horty et al.'s reasoner.

The conclusion that *Nixon* is anti-military is contentious. It could be argued that this is a reasonable conclusion; the graph contains no direct information about Republicans being pro-military to contradict the anti-military conclusion, and also the conclusion can be hedged by explicit mention of the founding class Quaker. Nevertheless, it could also be argued that since the graph is ambiguous about *Nixon* being a pacifist it should also be ambiguous about *Nixon* being anti-military, as this conclusion is a consequence of him being a pacifist. This type of argument was put forward by Touretzky et al. in support of a mechanism for propagating ambiguity. The consequences of trying to propagate ambiguity so that the graph becomes ambiguous about *Nixon* being anti-military will be discussed later, in section 6.9.

6.8.2 Upwards vs. Downwards Reasoners

The second choice presented by Touretzky et al. is the choice between upwards and downwards reasoning. This reasoner, like Horty et al.'s reasoner, can be considered to be an upwards reasoner. That is, it starts at the individual in question and moves up the inheritance net, deriving properties of that individual along the way. This, as noted by Touretzky et al., can be viewed as being similar to constructing proofs sequences in a logic.

The downward credulous reasoner of Touretzky [77] exhibits a phenomenon called coupling. In the homogeneous version of the graph displayed in figure 7[4] Touretzky's reasoner treats the nodes A and B identically. That is, even though the graph is ambiguous and generates multiple extensions, in any extension if the property E is inherited by members of B then E is also inherited by members of A. Similarly, if E is not inherited by the B's it is not inherited by the A's. There is no extension generated in which the conclusion about B's being E's is different from the conclusion about A's being E's. A and B can be said to be coupled. Upwards credulous

---

[4] **In figure 7 and 8 individuals are omitted—all of the nodes are property nodes. The individuals, i.e., the constant nodes, can be attached to any of the property nodes in the graph.**

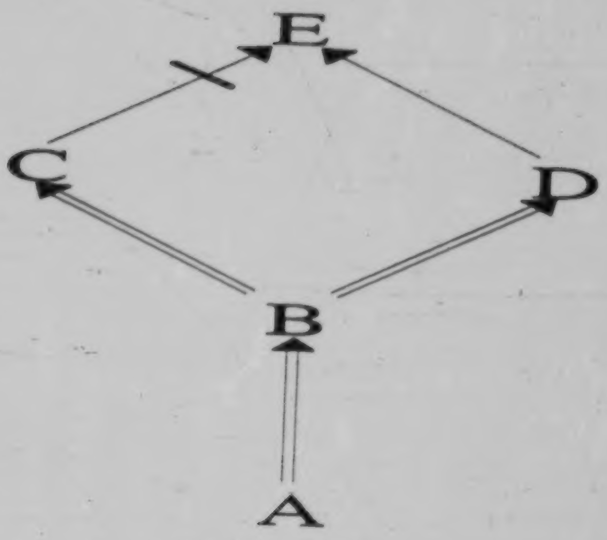

Figure 7: Nodes *A* and *B* are coupled

reasoners, on the other hand, do not exhibit coupling; e.g., in some of the extensions they generate, *A*'s may be *E*'s while *B*'s are not.

Even through the reasoner presented here is an upwards reasoner it also displays a form of coupling. Unlike Horty et al.'s skeptical reasoner which draws no conclusions about the *E*-ness of *A*'s or *B*'s, this reasoner concludes that the graph is ambiguous about *A*'s and *B*'s being *E*'s. That is, the conclusion about *A*'s and *B*'s is identical. The difference is: multiple extensions are not generated.

Another difference between upwards and downwards reasoners is that upwards reasoners display opportunism. In the homogeneous version of the graph in figure 8, Horty et al.'s skeptical upwards reasoner would conclude

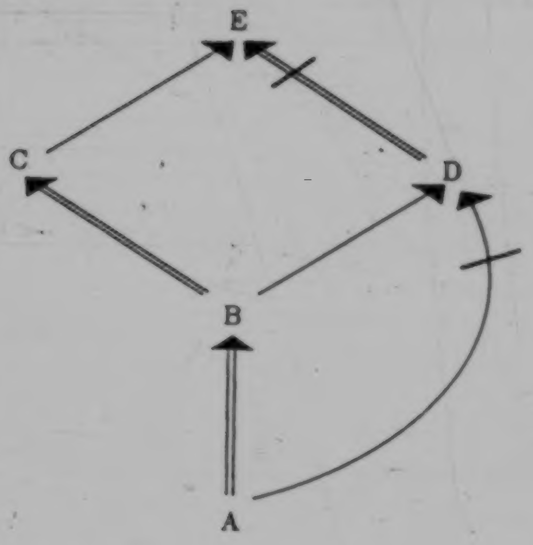

Figure 8: Opportunism

that *A*'s are *E*'s even through it draws no conclusions about *B*'s being *E*'s. Touretzky's downwards reasoner, however, can only conclude that *A*'s are *E*'s if it also concludes that *B*'s are *E*'s. That is, the extension in which *A*'s are *E*'s also contains the conclusion that *B*'s are *E*'s, while the other extension, in which *B*'s are not *E*'s, contains no conclusions about *A*'s being *E*'s.

This reasoner displays a form of opportunism in certain heterogeneous topologies. In particular, in the graph displayed in figure 8 the reasoner concludes that *A*'s are probably *E*'s founded on them being *C*'s, while at the same time it concludes that *B*'s are probably not *E*'s founded on them being *B*'s. This latter conclusion is different from the conclusion drawn by Horty et al.'s skeptical reasoner. Hence, the type of opportunism displayed by this reasoner is different and not directly comparable. (Note, the negative probable path from A to E founded on B is preempted, since the positive subpath from A to D is preempted.)

6.8.3 On-Path vs. Off-Path Preemption

The third choice is between on-path and off-path preemption. The technical details of the difference are complex, and not really relevant to this discussion. The difference can be amply explained by examining the graph in figure 9a. Touretzky's credulous reasoner, which uses on-path preemption, concludes that *Clyde's* grayness is ambiguous in the homogeneous version of this graph. Sandewall [69], on the other hand, claims that this graph should support the conclusion that *Clyde* is gray unambiguously and suggests the use of off-path preemption.

This reasoner behaves like an off-path preemptor. For the net in figure 9a it concludes that *Clyde* is probably not gray based on him being a Royal elephant. The probable positive path to *Gray* though *African elephant* is founded on *Clyde* being an elephant, and thus is preempted by the negative path.

It can be seen that this graph contains no information about the grayness of African elephants. African elephants inherit their grayness from

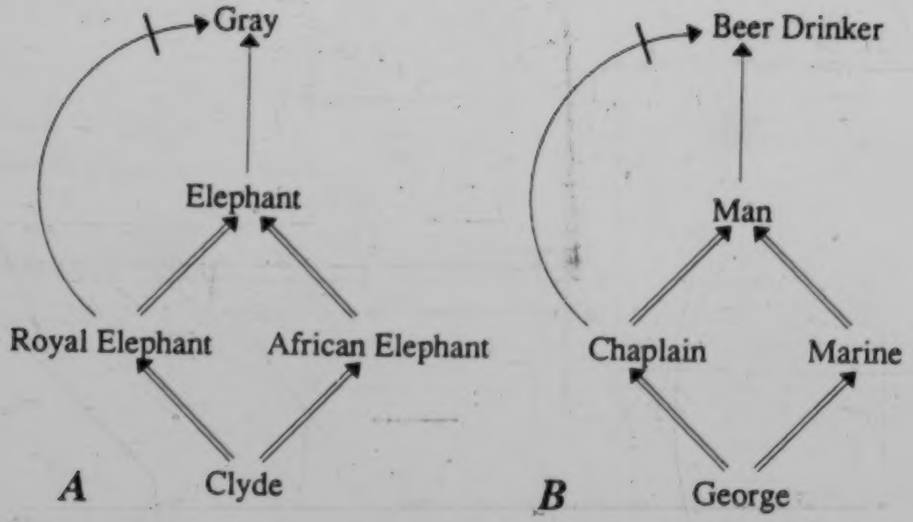

Figure 9: Two examples of On-path vs. Off-path preemption

the superclass elephant; thus, as Sandewall argued, the fact that *Clyde* is a Royal elephant, which are known to be normally non-gray, should override the normal grayness of average elephants. If we have specific knowledge about the grayness of African elephants, then this knowledge could be encoded in the graph through an explicit probable link from *African elephant* to *Gray*. In this case there would also be a non-preempted positive probable path from *Clyde* to *Gray*, and the graph would be ambiguous about *Clyde's* grayness.

The same reasoning applies to the graph in figure 9b. The homogeneous version of this graph was cited by Touretzky et al. as being a possible counter-example to Sandewall's argument. Touretzky et al. argue that *George's* beer drinking habits should be ambiguous in this graph, because, although he is a chaplain he is also a marine. They do, however, acknowledge that possibly the node *Marine* should have its own link to *Beer Drinker*, if we have explicit knowledge about the beer drinking habits of marines. When the knowledge in this graph is examined, it can be seen that the graph contains no information about the beer drinking properties of marines. Hence, with these semantics there is no choice but to add an explicit probable link from *Marine* to *Beer Drinker*, if marines are known to be heavy beer drinkers. If this is done the graph will become ambiguous about *George's* beer drinking habits.

6.9 Extending the Graph's Expressiveness

If the graph notation was extended to approach the expressiveness of the underlying logic, the complexity of finding paths in the graph would approach the complexity of theorem proving in the underlying logic. Path construction in the system presented here is easy because the graph can only express a simple set of formulas.

For example, consider the situation which would occur if conjunctive classes were allowed in the graph. In order for path preemption to continue to work properly, each conjunctive class would require an IS-A link to each of its constituents. For example, if the class $Republican \wedge Quaker$ was added to the Nixon diamond, probable paths from this conjunctive class should preempt probable paths from either *Republican* or *Quaker*. Preemption would only be possible if there was an IS-A link from this conjunctive class to both *Republican* and *Quaker*. Furthermore, individuals who were known to be both Republicans and Quakers would have to have an IS-A link to the conjunctive class. That is, *Nixon* would have to have a link to *Republican* ∧ *Quaker* in order for preemption to function correctly.

What is happening here is that deductive consequences are being added to the graph. For example, *Nixon* is known to be a *Republican* and is also known to be a *Quaker*; so, the deductive consequence that *Nixon* is a *Republican* ∧ *Quaker* must also be added to the graph. Such deductive consequences increase the number of paths in the graph exponentially; thus, the time required to check for path preemption will also grow.

Although limiting the graph to primitive classes yields an efficient reasoning system, it also poses certain problems. For example, propagating ambiguity in a sound manner is difficult. One simple means of propagating ambiguity is simply to make all paths which have an ambiguous subpath, ambiguous. However, propagating ambiguity in this manner would have the consequence of making the conclusion that *Nixon* is politically motivated in figure 4 ambiguous, as both paths to *Politically Motivated* have ambiguous subpaths.

The problem here is that although there is ambiguity about *Nixon* being

Hawk and about him being a ***Dove***, there is no ambiguity about him being a *Hawk* ∨ *Dove*. Thus, there should be no ambiguity about him being ***Politically Motivated***, as this conclusion comes from him being a *Hawk* ∨ *Dove*. The fact that the graph is incapable of representing disjunctive classes like *Hawk* ∨ *Dove* makes it difficult to propagate ambiguity in a sound manner—sometimes two ambiguities will cancel, sometimes they will not.

These problems point out that although efficient special purpose structured reasoners have an important role to play in a reasoning system, more general less efficient reasoners, like theorem provers, will also be required to handle the difficult cases.

Chapter 7

Conclusion

7.1 What has Been Accomplished

This thesis has dealt with the problem of representing and reasoning with probabilistic knowledge. It was motivated by a desire to include probabilistic knowledge in a general knowledge base and to extend the deductive inference mechanisms of first logic to mechanisms which could be used with probabilistic knowledge.

This led to an exploration of ways of mixing first order logic with probabilities. Viewing probabilities as a generalization of 0–1 truth values is a method which has been used extensively, both in philosophy and in AI. This approach attaches to sentences of first order logic graded truth values in the range 0–1, instead of just the endpoint 0, 1 values. This can be accomplished by positing a probability distribution over the Lindenbaum-Tarski algebra formed by the equivalence classes of provably equivalent sentences. Equivalently, a probability distribution can be placed over a set of possible worlds, where each possible world in this set is a unique assignment of truth values to the sentences of the logic. This approach is, however, inadequate for the expression of probabilistic knowledge which takes the form of statistical assertions.

Statistical assertions are statements of empirical probability which assert something about the world. Although empirical probability statements are not the only kind of probability assertion that we wish to deal with, they are an essential subset of the probabilistic knowledge possessed by any rational agent. Their importance lies in the fact that they can in principle be 'learned' through accumulated experience with the environment.

In order to capture empirical probabilities a new mixture of probability and logic was needed. A natural alternative candidate over which a probability distribution could be placed is the domain of discourse. The existence of the negation and conjunction operators in first order logic ensures that the collection of sets defined by the formulas of the logic form a field of sets, the minimal structure over which a probability distribution can be placed. The development of this intuition led to the major contribution of the thesis, the logic **Lp**.

Lp is an extension of first order logic: the model structure has been extended to include a probability function over the domain of discourse, and the syntax extended to allow the formation of probability terms which can be used to express statements of empirical probability. Another innovation was to embed a totally ordered field of numbers in the semantic model. This allows access in the syntax to the values of the probability function. By making the logic two sorted, the probability terms can participate in the formation of complex sentences which allow the expression of a very general set of probability assertions, including many assertions which make no mention of particular numbers.

Although this logic solves the problem of representing statements of empirical probability, probabilistic degrees of belief assigned to sentences of the logic remain outside its scope. In contrast, the existing approach of a probability distribution over a set of possible worlds is capable of expressing such degrees of belief.

The second contribution of the thesis is the development of a mechanism of belief formation. This mechanism is capable of using the empirical probabilities expressed in **Lp** to generate probabilistic degrees of belief in a large class of logical sentences. These degrees of belief offer a major advantage over the purely subjective possible worlds approach: they come from empirical data, i.e., they are founded on knowledge about the world. This goes a long way towards answering the common question "Where do the probabilities come from," especially in view of the fact that the underlying logic is capable of expressing vague, non-numeric assertions of probability, i.e., numbers are not required. This approach also has the advantage that the empirical knowledge can in principle be accumulated through experience.

Another feature of these degrees of belief is that they display nonmonotonic behavior; thus, the mechanism offers an alternative formalism for reasoning with defaults which have a statistical interpretation. This feature has been used to give a treatment of multiple inheritance hierarchies with exceptions, a problem which is beyond the monotonic capabilities of strictly deductive inference. The formalism gives a natural statistical interpretation to the defeasible links in the hierarchy, and can, in its full generality, deal with composite classes and n-place predicates. The treatment of inheritance hierarchies demonstrates some of the generality of the formalism.

7.2 Future Research

There are three important extensions to this work which I feel are promising areas for future research: extending the mechanism of belief formation, developing a theory of diagnosis from statistical principles, and learning from experience. I will conclude this thesis with a more detailed discussion of these areas of research.

7.2.1 Extensions to Belief Formation

Degrees of Belief

One obvious weakness of the mechanism of belief formation is that the degrees of belief lie outside of the logical formalism. One possible solution is to combine a possible worlds approach with the approach of **Lp** to get a logic capable of expressing both empirical probabilities as well as probabilities attached to sentences. In this approach each possible world would be an **Lp**-structure, with a probability distribution over its domain of discourse. In addition there would be a probability distribution over the set of **Lp**-structures. This latter probability distribution would give a probability assignment to the **Lp** sentences—the probability of a sentence would be the measure of the set of worlds in which it is true.

This combination would have the advantage of giving the degrees of belief an explicit semantics. Furthermore, it would allow the development of a proof theory for the degrees of belief, such a proof theory would allow reasoning with the degrees of belief directly, i.e., without having to returning to the level of **Lp**.

The difficulty lies in capturing the inductive principle in this extended formalism. The inductive principle expresses a reasonable relationship between the probabilities assigned to the **Lp** sentences and the empirical probabilities in the **Lp**-structure. One can give up this relationship only on pain of returning to unfounded subjective probabilities.

Further Theoretical Justifications for Belief Formation

When an agent's beliefs are simply a set of new facts deduced from his knowledge base using sound rules of inference, there is a simple relationship between the agent's beliefs and the state of the world. If the agent's factual knowledge is accurate then he will believe assertions which are true. This gives a good reason for claiming that the agent is rational. However, if the agent uses belief formation and **Lp** deduction to generate degrees of belief, there will be no such direct assurance that the agent is rational. This is because belief formation does not deduce a sentence α; instead it imparts a degree of belief to the sentence α. Hence, we are dealing with an agent who, after some deduction in his internal language, believes a sentence α to some degree. Characterizing the relationship between these degrees of belief and the state of the world is a much more difficult problem. All that can be hoped for is that if the agent accumulates sufficient information about the world and is able to devote sufficient computational resources to evaluating his degrees of belief, he will tend to assign high degrees of belief to statements which are true statements about the world and low degrees of belief to statements which are false. In essence, this would require showing that the mechanism of belief formation, and its associated inductive assumption of randomization, is in fact a rational mechanism in the long run.

It has already been shown that if random selection is used the degrees of belief are justified by long term trends. However, a stronger justification would involve showing that an agent acting on these induced degrees of belief (i.e., using these degrees of belief as guides to action) would be better off, in the long run, than if he acted on any other degrees of belief. A possible approach to this difficult problem would be to show that in the context of certain statistical games against nature the inductive principle is optimal. It may be possible to show that if the game has certain rules (e.g., a lack of knowledge of nature's selection rule) then the inductive principle is the best strategy. A related project would be to justify the preference criterion by showing, for example, that a player who used it would do better than a player who did not.

Beliefs About Beliefs

Work on formal models of belief has been motivated by a desire to reason about other agents' beliefs, something which is important when reasoning about their actions. This involves being able to reason with iterated belief statements like "John believes that Mary believes he loves her." In order to deal with many agents each having their own beliefs, including beliefs about each others beliefs, modal logics are required. Modal logics can represent the different world views possessed by distinct agents and can reason with the relationships between those different views. An interesting project would be to extend the formalism developed in this thesis to a modal logic that would be capable of reasoning about a collection of interacting agents. One attractive feature of this formalism lies in its ability to generate beliefs from incomplete information. When one considers how an agent could generate beliefs about another agent's beliefs it is seen that such beliefs must be based on incomplete knowledge. For example, how could John arrive at the conclusion that Mary believes he loves her? John does not have direct access to Mary's beliefs; instead, he must base his beliefs about Mary's beliefs on Mary's actions as well as on knowledge about the world, e.g., knowledge of how people act when they think someone is in love with them. Such knowledge has a distinctly statistical flavour; i.e., people "usually" act in a certain manner when they hold certain beliefs, but not always.

Universal Sentences

Another extension to the mechanism of belief formation would be to deal with universally quantified sentences. Belief formation can only say trivial things about universal sentences which contain no object constants, e.g., "$\forall x\, P(x) \rightarrow Q(x)$." If the sentence or its negation is deducible from the knowledge base, then belief formation will assign a degree of belief to the sentence equal to one or zero. Otherwise, nothing can be deduced about the sentence's degree of belief except for the trivial fact that it is in the range 0–1. The reason for this is that there are no individuals in the sentence that can be randomized. Hence, some other inductive principle must be found to deal with universals.

Universals have always posed a problem for induction, due to Hempel's paradox of confirmation. One example of the paradox is as follows. Suppose that we are interested in the universal sentence "All ravens are black," then a logically equivalent sentence is "All non-black object are not ravens." If a universal sentence is to be confirmed by observing a large number of ground instances, then observing a large number of snow flakes in a snow storm would give a lot of evidence for the second sentence, and thus would also be confirming evidence for the assertion about ravens. This is obviously an unnatural conclusion.

One approach to this problem would be to assert that universals are never induced from experience. Instead what is induced are conditional probabilities, e.g., $[Black(x)|Raven(x)]_x$. The confirmation of conditionals does not suffer from this paradox, since only observations of ravens count. However, this does not seem to be the case: assertions of universal laws can be found in many areas, e.g., in physics. These universals have been induced from observations, although in many cases not solely from large sets of ground instances. It has been noted (Russell [67]) that the induction of universals requires additional knowledge, besides a set of confirming ground instances. For example, in the domain of physics an important piece of knowledge used in the induction of universals is the fact that physical laws generally are universal. Another example is that of the black ravens. If it is known that many species of birds do exhibit uniform colouration, it would be reasonable to induce a high degree of belief in the universal if all of the ravens that have been observed have been black. It is possible that these higher level generalizations may be statistical generalizations expressible in **Lp**, and that a mechanism could be developed to assign degrees of belief to lower level universals.

7.2.2 Diagnosis from Statistical Principles

The results on inheritance hierarchies presented in this thesis have demonstrated how certain types of defaults can be profitably represented by high conditional probabilities. Simple conditional probabilities are not sufficient to encode the causal relationships present in domains where diagnosis is possible. The following example, due to Len Schubert, makes this clear:

- Consider two sets of statistical assertions about the diseases D_1 and D_2 and their relation to the symptom S. The first set of assertions is

$$\begin{array}{ll} [D_1(x)]_x = 0.1, & [S(x)|D_1(x)]_x = 0.9 \\ [D_2(x)]_x = 0.9 & [S(x)|D_2(x)]_x = 0.1, \end{array}$$

while the second set is

$$[D_1(x)]_x = 0.9, \qquad [S(x)|D_1(x)]_x = 0.1$$

$$[D_2(x)]_x = 0.1 \qquad [S(x)|D_2(x)]_x = 0.9.$$

Clearly, the symptom S has a different significance in these two cases. In the first case, S usually accompanies D_1 but rarely accompanies D_2. In the second case, S is, instead, correlated with D_2. Intuitively, given the first set of assertions one would think that S provides good evidence for (tends to confirm) D_1, while given the second set one would think that S provides evidence for D_2. Yet, as can be easily verified, for both sets of assertions the inverse conditionals are equal: from the first set of assertions it can be deduced that $[D_1(x)|S(x)]_x = [D_2(x)|S(x)]_x = 0.5$, while the second set derives $[D_1(x)|S(x)]_x = [D_2(x)|S(x)]_x = 0.5$.

This example suggests that what is important in measuring evidential support is not just the conditional probabilities, but is instead their magnitudes relative to the unconditional probabilities. Measures of causal support can be expressed by ratios of probability terms. It is easy to see that such ratios are expressible in **Lp**, and thus the causal information used in diagnosis can be represented in the formalism.

Of course, the representation of causal information is just a start. There are a number of other problems which would have to be solved before a workable diagnosis system could be developed. First, the mechanism of belief formation (which would be used to reason about particular cases using the background statistical knowledge expressed in **Lp**) has only been specified in a very general manner. A more specific control structure must be developed for an effective computer diagnosis system. One possibility is to use the causal networks developed by Pearl [55] as a special purpose reasoning structure. However, what is really required for a diagnosis system with a wide range of coverage is the ability to dynamically configure such a reasoning structure depending on what the system is reasoning about. For this, backwards reasoning from effects to causes (e.g., Morgan [51], Poole et al. [61]) may be very useful. A set of possible causes which explain the observed symptoms (explain in a more general sense of making probable, instead of logically entailing) could be generated by such methods and then structured into a causal net. Once the causal net was constructed probabilistic analysis could be performed.

One useful feature of this formalism is that it is based on probabilities; hence, it leads naturally into decision theory. Much of diagnosis deals with tradeoffs. For example, should more tests be performed or should a treatment be prescribed which will deal with the most probable disorder? These tradeoffs can be approached rationally from the standpoint of decision theory. For example, given knowledge of a treatment's side effects the above question could be answered by considering the expected utility of the treatment. One would expect that any powerful theory of diagnosis would incorporate notions from decision theory.

7.2.3 Learning, or Induction, from Ground Facts

One of the most attractive features of this work is that the statistical knowledge expressed in **Lp** could be accumulated through experience. To specify exactly how this could be done would be a major piece of research. There are a large number of methods in statistics which may be applicable. These methods use data from the world, i.e., samples, to assign degrees of credence to various statistical assertions, i.e., hypotheses. The most difficult problem, however, is the problem of acceptance.

As it stands now, it has been assumed that there is some external agent, e.g., a domain expert, who has decided what the accepted knowledge is and has encoded this knowledge as a set of **Lp** sentences. From this already accepted knowledge, uncertain beliefs are generated through belief formation. Hence, the problem of acceptance has not arisen. If we wish to develop automated mechanisms which can accumulate evidence for various assertions and then accept those assertions into the knowledge base when sufficient evidence has been accumulated, then we must face the lottery paradox. This paradox, due to Kyburg [39], is as follows:

- A large number of people buy tickets to a lottery having a single winner. The probability that the i-th person wins the lottery is very small and can be made arbitrarily small as the number of tickets increase. If an assertion is to be accepted as true when it has a very

> high probability, then every statement "person i is not the winner" would be accepted. But, the collection of all these statements is inconsistent with the fact that one person will definitely be the winner.

Some recent work by Kyburg [36] has produced what seems to be a workable solution to this problem. Levi and Morgenbesser [43] have previously pointed out that to a certain extent there are no such things as fully accepted beliefs, to quote:

> "for any contingent proposition p on which action can be taken, there is a least one objective relative to which a nonsuicidal, rational agent would refuse to act as if p were true. Consider, for example, the following gamble on the truth value of p: If the agent bets on p and p is true, he wins some paltry prize, and if p is false he forfeits his life. However, if he bets on $\neg p$, he stands to win or lose some minor stake. ... (Hence) ... the agent could not rationally and sincerely believe p where p is contingent."

This argues that the acceptance of beliefs involves some notion of utility. Kyburg constructs a system in which beliefs are tagged by their odds, i.e., their probability divided by one minus their probability. For actions where the ratio of risk to reward is less that the tagged odds, the agent will act as if he has fully accepted the belief. For example, if you park your car in a parking lot you may accept the belief that it will still be there when you return later. In this case if the odds of the car being there are high, say 1000:1, then if the ratio of risk to reward is less than 1000:1, you are justified in accepting the belief. For example, it may cost you $20 to take a cab home versus a benefit of, say, $1 if the car is there. Since the odds of the conjunction of two statements must be lower, two statements may be accepted at some level of risk to reward while their conjunction is not; thus, the lottery paradox is avoided. By adapting these ideas it may be possible to inductively accept assertions into the knowledge base, perhaps by partitioning the knowledge base into levels of acceptance indexed by risk to reward ratios.

Bibliography

[1] Lennard Åqvist, Jaap Hoepelman, and Christian Rohrer. Adverbs of frequency. In Christian Rohrer, editor, *Time, Tense and Quantifiers: Proceedings of the Stuttgart Conference on the Logic of Tense and Quantification*, M. Niemeyer, Tubingen, 1980.

[2] Romas Aleliunas. *Models of Reasoning Based on Formal Deductive Probability Theories*. Technical Report, University of Waterloo, 1986.

[3] John Bell and Moshé Machover. *A Course in Mathematical Logic*. Elsevier, Netherlands, 1977.

[4] D. Bobrow (ed.). Special issue on nonmonotonic reasoning. *Artificial Intelligence*, 13, 1980.

[5] Ronald J. Brachman. I lied about the trees. *AI Magazine*, 6(3):80–93, 1985.

[6] Alan Bundy. Incidence calculus: a mechanism for probabilistic reasoning. *Journal of Automated Reasoning*, 1:263–283, 1985.

[7] G. Carlson. *Reference to Kinds in English*. PhD thesis, University of Massachusetts, 1977. unpublished—available from Indiana University Linguistics Club.

[8] R. Carnap and R.C. Jeffrey. *Studies in Inductive Probability*. Univ. of California Press, Berkeley and Los Angeles, CA, 1971.

[9] Rudolf Carnap. *Logical Foundations of Probability*. University of Chicago Press, 1962.

[10] Kai Lai Chung. *A Course in Probability Theory*. Academic Press, New York, 1974.

[11] R. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14(1):1–13, January-February 1946.

[12] Martin Davis. The mathematics of non-monotonic reasoning. *Artificial Intelligence*, 13, 1980.

[13] B. de Finetti. *Probability, Induction and Statistics*. Wiley, New York, 1972.

[14] Bruno De Finetti. Foresight: its logical laws, its subjective sources. In Kyburg and Smokler, editors, ***Studies in Subjective Probability***, John Wiley and Sons, New York, 1964.

[15] Johannes de Haan and L.K. Schubert. Inference in a topically organized semantic net. In *AAAI-86*, pages 334–338, 1986.

[16] James P. Delgrande. An approach to default reasoning based on a first-order conditional logic. In *AAAI-87*, pages 340–335, 1987.

[17] Richard O. Duda, Peter E. Hart, and Nils J. Nilsson. Subjective bayesian methods for rule-based inference systems. In Bonnie Lynn Webber and Nils J. Nilsson, editors, ***Readings in Artificial Intelligence***, pages 192–199, Morgan Kaufmann, 1981.

[18] David W. Etherington. Formalizing nonmonotonic reasoning systems. *Artificial Intelligence*, 31:41–85, 1987.

[19] David W. Etherington. More on inheritance hierarchies with exceptions. In *AAAI-87*, pages 352–357, 1987.

[20] Ronald Fagin, Joseph Y. Halpern, and Nimrod Megiddo. ***A Logic For Reasoning About Probabilities.*** Technical Report RJ 6190 4/88, IBM Research, Almaden Research Center, 650 Harry Road, San Jose, California, 95120–6099, 1988.

[21] H. Field. Logic, meaning, and conceptual role. *Journal of Philosophy*, 77:374–409, 1977.

[22] H. Gaifman. Concerning measures in first order calculi. *Israel Journal of Mathematics*, 2:1–18, 1964.

[23] Hector Geffner and Judea Pearl. *Sound Defeasible Inference*. Technical Report CSD870058, Cognitive Systems Laboratory, U.C.L.A., Los Angeles, CA. 90024–1596, U.S.A., 1987.

[24] Benjamin N. Grosof. An inequality paradigm for probabilistic knowledge. In L.N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 259–275, North-Holland, 1986.

[25] Benjamin N. Grosof. Non-monotonicity in probabilistic reasoning. In *Proceedings of the AAAI/RCA Workshop on Uncertainty and Probability in Artificial Intelligence*, pages 91–98, 1986.

[26] David Heckerman. Probabilistic interpretations for MYCIN's certainty factors. In L.N. Kanal and J.F. Lemmer, editors, *Uncertainty in Artificial Intelligence*, pages 167–196, North-Holland, 1986.

[27] David E. Heckerman and Eric J. Horvitz. On the expressiveness of rule-based systems for reasoning with uncertainty. In *AAAI-87*, pages 121–126, 1987.

[28] Carl G. Hempel. Deductive-nomological vs. statistical explanation. In Herbert Feigland and Grover Maxwell, editors, *Minnesota Studies in the Philosophy of Science Vol III*, pages 98–169, University of Minnesota Press, Minneapolis, 1962.

[29] J. Hintikka. A two-dimensional continuum of inductive methods. In J. Hintikka and P. Suppes, editors, *Aspects of Inductive Logic*, Amsterdam, 1966.

[30] John F. Horty, Richmond H. Thomason, and David S. Touretzky. A skeptical theory of inheritance in nonmonotonic semantic networks. In *AAAI-87*, pages 358–363, 1987.

[31] E.J. Horvitz, D.E. Heckerman, and C.P. Langlotz. A framework for comparing alternative formalisms for plausible reasoning. In *AAAI-86*, pages 210–214, 1986.

[32] R.W. Johnson. Independence and bayesian updating methods. *Artificial Intelligence*, 29:217–222, 1986.

[33] H.J. Keisler. Probability quantifiers. In J. Barwise and S. Feferman, editors, *Model Theoretic Logics*, chapter XIV, Springer, N.Y., 1985.

[34] John G. Kemeny. Fair bets and inductive probabilities. *The Journal of Symbolic Logic*, 20–3:263–273, September 1955.

[35] A. Kolmogoroff. *Foundations of the Theory of Probability*. Chelsea Publishing Company, New York, 1950.

[36] Henry E. Kyburg, Jr. Full beliefs. *submitted to Theory and Decision*, 1987.

[37] Henry E. Kyburg, Jr. *The Logical Foundations of Statistical Inference*. D. Reidel, 1974.

[38] Henry E. Kyburg, Jr. *Probability and Inductive Logic*. Macmillan, 1970.

[39] Henry E. Kyburg, Jr. *Probability and the Logic of Rational Belief*. Wesleyan University Press, Middletown, Connecticut, 1961.

[40] Henry E. Kyburg, Jr. The reference class. *Philosophy of Science*, 50(3):374–397, September 1983.

[41] H. LeBlanc. Alternatives to standard first-order semantics. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic. Vol II*, pages 225–258, Reidel, Holland, 1983.

[42] R. Sherman Lehman. On confirmation and rational betting. *The Journal of Symbolic Logic*, 20–3:251–262, September 1955.

[43] Levi and Mogenbesser. Belief and disposition. *American Philosophical Quarterly*, 221–232, 1964.

[44] Vladimir Lifschitz. *Pointwise Circumscription.* Technical Report, Stanford University, Computer Science Department, 1986.

[45] S. MacLane and G. Birkhoff. *Algebra.* Macmillan, New York, 1968.

[46] John McCarthy. Applications of circumscription to formalizing common-sense knowledge. *Artificial Intelligence*, 28:86–116, 1986.

[47] John McCarthy. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence*, 13:27–39, 1980.

[48] John McCarthy and Patrick Hayes. Some philosophical problems from the standpoint of artificial intelligence. In Bonnie Lynn Webber and Nils J. Nilsson, editors, *Readings in Artificial Intelligence*, pages 431–450, Morgan Kaufmann, 1981.

[49] Drew McDermott and Jon Doyle. Non-monotonic logic I. *Artificial Intelligence*, 13:41–72, 1980.

[50] C. Morgan. Weak conditional comparative probability as a formal semantic theory. *Zeit. fur Math. Log.*, 30:199–212, 1984.

[51] Charles G. Morgan. Hypothesis generation by machine. *Artificial Intelligence*, 2:179–187, 1971.

[52] J. Neyman. *First Course in Probability and Statistics.* New York, 1950.

[53] Nils J. Nilsson. Probabilistic logic. *Artificial Intelligence*, 28:71–87, 1986.

[54] Jane Terry Nutter. Uncertainty and probability. In *Proceedings of the 10th IJCAI*, pages 373–379, 1987.

[55] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artificial Intelligence*, 29:241–288, 1986.

[56] Judea Pearl. On the logic of probabilistic dependencies. In *AAAI-86*, pages 339–343, 1986.

[57] Judea Pearl. *Probabilistic Semantics for Inheritance Hierarchies with Exceptions.* Technical Report CSD870052, Cognitive Systems Laboratory, U.C.L.A., Los Angeles, CA. 90024–1596, U.S.A., 1987.

[58] Judea Pearl and Paz Azaria. On the logic of representing dependencies by graphs. In *Proceedings of Sixth Canadian Artificial Intelligence Conference*, pages 94–98, 1986.

[59] Judea Pearl and Thomas Verma. The logic of representing dependencies by directed graphs. In *AAAI-87*, pages 374–379, 1987.

[60] Francis Jeffry Pelletier and Lehnart K. Schubert. *Three Papers on the Logical Form of Mass Terms, Generics, Base Plurals, and Habit-*

uals. Technical Report 87–3, Department of Computing Science, The University of Alberta, Edmonon, Alberta, Canada. T6G–2H1, 1987.

[61] David Poole, Randy Goebel, and Romas Aleluinas. Theorist: a logical reasoning system for defaults and diagnosis. In N.J. Cercone and G. McCalla, editors, *The Knowledge Frontier: Essays in the Representation of Knowledge*, pages 331–352, Springer-Verlag, New York, 1987.

[62] David L. Poole. *Default Reasoning and Diagnosis as Theory Formation*. Technical Report CS–86–08, Department of Computing Science, University of Waterloo, Waterloo, Ontario, Canada. N2L–3G1, 1986.

[63] K.R. Popper. The propensity interpretation of probability. *British Journal for the Philosophy of Science*, 10:25–42, 1959.

[64] Hans Reichenbach. *Theory of Probability*. University of California Press, Berkeley and Los Angeles, CA., 1949.

[65] Raymond Reiter. A logic for default reasoning. *Artificial Intelligence*, 13, 1980.

[66] J.B. Rosser and A.R. Turquette. *Many-Valued Logic*. North-Holland, 1952.

[67] Stuart J. Russell. Preliminary steps towards the automation of induction. In *AAAI-86*, pages 477–484, 1986.

[68] Wesley Salmon. *The Foundations of Scientific Inference.* University of Pittsburgh Press, Pittsburgh, 1967.

[69] Erik Sandewall. Nonmonotonic inference rules for multiple inheritance with exceptions. *Proceedings of the IEEE*, 74(10):1345–1353, October 1986.

[70] Leonard J. Savage. *The Foundations of Statistics.* Dover, New York, 1964.

[71] Ross D. Schachter and David Heckerman. A backwards view for assessment. *AI Magazine*, 6(3):55–62, 1987.

[72] Mark J. Schervish, Teddy Seidenfeld, and Joseph B. Kadane. The extent of non-conglomerability of finitely additive probabilities. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66:205–226, 1984.

[73] L. K. Schubert, M. A. Papalaskaris, and J. Taugher. Accelerating deductive inference: special methods for taxonomies, colours, and times. In N.J. Cercone and G. McCalla, editors, *The Knowledge Frontier: Essays in the Representation of Knowledge*, pages 187–220, Springer-Verlag, New York, 1987.

[74] Abner Shimony. Coherence and the axioms of confirmation. *The Journal of Symbolic Logic*, 20-1:1–28, March 1955.

[75] Edward H. Shortliffe and Bruce G. Buchanan. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23:351–379, 1975.

[76] A. Tarski. *A Desision Method for Elementary Algebra and Geometry, 2nd Edition.* University of California Press, 1951.

[77] David S. Touretzky. *The Mathematics of Inheritance Systems. Research Notes in Artificial Intelligence*, Pitman, London, 1986.

[78] David S. Touretzky, John F. Horty, and Richmond H. Thomason. A clash of intuitions: the current state of nonmonotonic multiple inheritance systems. In *IJCAI-87*, pages 476–482, 1987.

[79] B. van Fraassen. Probabilistic semantics objectified. *Journal of Philosophic Logic*, 10:371–394, 1981.

[80] John Venn. *The Logic of Chance.* Chelsea Publishing Company, London, 1866.

[81] Richard von Mises. *Probability, Statistics, and Truth.* Dover Publications, Inc., New York, 1957.